

Copyright  
by  
Dacia Leon  
2018

**The Dissertation Committee for Dacia Leon Certifies that this is the approved  
version of the following Dissertation:**

**Directed evolution of antimutator *E. coli***

**Committee:**

---

Jeffrey Barrick, Supervisor

---

Arlen Johnson

---

Edward Marcotte

---

Tanya Paull

---

Claus Wilke

**Directed evolution of antimutator *E. coli***

**by**

**Dacia Leon**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

**December 2018**

## **Dedication**

This work is dedicated to my family – my mother, father, stepparents, and Jared. And tacos, my fuel for science.

## **Acknowledgements**

I love science. However, sometimes (all the time) it proves to be challenging and a bit discouraging (banging head against idea/experiment/data). I've been lucky to be surrounded by incredibly intelligent, positive, and inspiring people who make science exciting, despite all of the hurdles. Namely, I'd like to thank my wonderful advisor Jeffrey Barrick. I feel as if I've won the advisor lottery with Jeff as my mentor – he is supportive, enthusiastic, encouraging, funny!, and always asks for the best. I want to thank Claus and the members of the Wilke lab for providing me with a second scientific home. I would also like to thank the rest of my committee members, Arlen Johnson, Edward Marcotte, and Tanya Paull, for their guidance and helpful advice in staying on course.

I am so grateful for the close community of members in the Barrick lab. I'd like to particularly thank some of my past and present labmates for being my friends and collaborating with me – Colin Brown, Simon D'Alton, Dan Deatherage, Kyle Dugan, Peng Geng, Mike Hammerling, Jenna McGuffey, Julie Perreau, Erik Quandt, and Brian Renda.

Lastly, and most importantly, I'm indebted to my family for their unconditional support and love, I owe my accomplishments to them. And to Jared, for always being a rat.

## Abstract

### Directed evolution of antimutator *E. coli*

Dacia Leon, Ph.D.

The University of Texas at Austin, 2018

Supervisor: Jeffrey E. Barrick

Biological systems are essential tools for addressing societal challenges. There have been several successes in this field, however, a strong hindrance lies in the ephemeral nature of these systems – cells are tiny factories that evolve. Evolution poses a problem because when a desired function is encoded into the DNA of the host organism, the host uses its own resources to perform the function and there is likely no associated fitness benefit. Therefore, there is strong selection for inactivation of the function due to the metabolic load imposed on cellular resources. One way to address this problem is to engineer evolutionary stability by lowering a host organism's basal mutation rate and concomitantly reducing the probability that an encoded function will become mutated.

In **Chapter 1** of this dissertation, I discuss the nature of the metabolic cost associated with engineering biology and mechanisms by which host adaptation occurs. I also explore cellular pathways involved in genetic stability and examine previously characterized antimutators. **Chapter 2** describes the first iteration of a directed evolution method used to engineer antimutators in *Escherichia coli*, Periodic Reselection for Evolutionarily Reliable Variants (PResERV). In this first PResERV experiment, I observe that the antimutator phenotype is due to mutations in genes involved DNA replication and

RNA metabolism (*polA*, *polB*, and *rne*). In **Chapter 3**, I perform the same PResERV experiment on a greater scale and characterize a series of antimutator strains. The causative alleles in many of these strains are in genes involved in the tricarboxylic acid cycle and electron transport chain (*sucD* and *sdhA*). These alleles are shown to reduce oxidative stress. **Chapter 4** demonstrates results from another PResERV experiment using a clean-genome *E. coli* strain, MDS42, as the host organism. In sum, this work shows the many mechanisms that lead to an antimutator phenotype, and these findings are used to build stable strains for reliable engineering of biology.

Finally, there are two appendices (**Appendix A** and **B**) which discuss my work in examining the evolutionary path to citrate utilization in Lenski's long-term evolution experiment (LTEE) and a do-it-yourself method for using gellan gum as an alternative to microbial agar media.

## Table of Contents

List of Tables .....	xiv
List of Figures .....	xv
Chapter 1: Introduction .....	1
Metabolic cost and adaptations associated with engineering biology .....	3
Cellular Pathways that contribute to genetic instability .....	5
Previously described antimutators .....	9
Chapter 2: Directed evolution of <i>Escherichia coli</i> with lower-than-natural plasmid mutation rates .....	13
Abstract .....	14
Introduction .....	15
Results .....	16
PResERV experiment with a ColE1 plasmid in <i>E. coli</i> .....	16
Mutations in PResERV strains .....	17
Evolutionary stability and mutation rates in PResERV and reconstructed strains .....	19
Plasmid copy number and GFP fluorescence in evolved strains .....	20
Discussion .....	23
Materials and Methods .....	36
Culture conditions .....	36
Strains and plasmids .....	36
UV mutagenesis .....	37
PResERV directed evolution procedure .....	37
Isolation of evolved cells and plasmids .....	38



GFP decay curves .....	39
Genome sequencing.....	39
Strain reconstruction.....	40
Mutation rate measurements .....	40
Plasmid copy number determination .....	42
Scaling of apparent mutation rates with plasmid copy number.....	43
Chapter 3: Reduced plasmid mutation rates in <i>Escherichia coli</i> evolve by mitigating the generation of reactive oxygen species through mutations that inactivate TCA cycle enzymes .....	56
Abstract .....	56
Introduction .....	57
Results.....	58
Directed evolution via PResERV .....	58
Evolved strains have reduced mutation rates.....	59
Most PResERV strains maintain wild-type plasmid copy number and growth rates.....	60
Mutations in the genomes of PResERV strains .....	61
Mapping the antimutator phenotype.....	62
Loss of function mutations in <i>sucD</i> and <i>sdhA</i> reduce the <i>E. coli</i> basal mutation rate .....	63
Further dissection of the antimutator roles of succinyl-CoA synthetase and succinate dehydrogenase.....	65
Antimutator alleles function to alleviate DNA damage induced by ROS ...	65
PResERV alleles reduce mutation rates in other strains .....	67
Total Antimutator Combination Organism (TACO).....	67

Discussion.....	68
Different mutational solutions to the same problem .....	69
How do loss-of-function antimutator alleles in <i>sucD</i> and <i>sdhA</i> result in decreased ROS? .....	70
<i>sucD</i> .....	70
<i>sdhA</i> .....	71
Rational engineering of antimutators by targeting ROS pathways .....	73
Materials and Methods.....	74
Culture conditions.....	74
Strains and plasmids .....	75
UV mutagenesis .....	75
PResERV.....	76
Isolation of clones with increased GFP stability .....	77
Mutation rate measurements .....	78
Genome sequencing.....	79
Absolute plasmid copy number .....	79
Growth rate measurements.....	80
Bacteriophage transduction.....	81
Strain construction for reversion mutants .....	82
Paraquat survival assay .....	82
Intracellular ROS assay .....	82
Chapter 4: PResERV antimutators evolved from a clean-genome strain, MDS42 .....	93
Introduction .....	93

Results.....	94
Rapid inactivation of a costly plasmid in <i>E. coli</i> .....	94
PResERV evolution of MDS42 .....	95
Genetic stability of MDS42 PResERV evolved clones .....	96
Mutation rates of evolved clones .....	96
Whole genome sequence analysis of MDS42 evolved clones.....	97
Future Directions.....	99
Materials and Methods.....	100
Culture conditions.....	100
Strains and plasmids .....	100
UV mutagenesis .....	101
PResERV directed evolution .....	101
GFP decay curves .....	102
Genome sequencing.....	103
Mutation rate measurements .....	104
Chapter 5: Conclusion and Future Directions.....	119
Appendix A: Innovation in an <i>E. coli</i> evolution experiment is contingent on maintaining adaptive potential until competition subsides.....	123
Abstract .....	124
Author Summary.....	125
Introduction .....	125
Results.....	128
Cit <sup>+</sup> was only slightly beneficial when it evolved in the LTEE.....	128

Development of $P_{rnk-citT}$ knock-in assay for potentiation.....	129
Cit <sup>+</sup> would have been modestly beneficial if it evolved in the LTEE ancestor .....	131
No evidence for ecological potentiation.....	132
Anti-potentiated strains evolved at intermediate time points .....	132
Mapping potentiation onto phylogeny .....	134
Cit <sup>+</sup> evolution in the context of competition with other beneficial mutations.....	136
Discussion.....	138
Materials and Methods.....	145
Media conditions and strains .....	145
$P_{rnk-citT}$ knock-in assay.....	145
Selection for spontaneous Ara <sup>+</sup> mutants.....	146
Relative fitness measurements .....	147
qRT-PCR measurement of <i>citT</i> expression.....	148
Genome sequencing and phylogenetic tree construction .....	150
Beneficial mutation fitness effect models .....	150
Appendix B: Bacterial production of gellan gum as a do-it-yourself alternative to agar	161
Abstract .....	162
Introduction .....	163
Methods and Results .....	163
I. Prepare DIY media for gellan gum production .....	163
II. Culture <i>S. paucimobilis</i> for gellan gum production.....	164
III. Pour gellan gum plates .....	165

Safety issues .....	165
Discussion.....	165
References .....	168

## **List of Tables**

Table 3.1. Candidate antimutator alleles identified via PResERV .....	92
Table 4.1. Mutations in MDS42 PResERV strains .....	113
Table A1. Genome sequencing of <i>E. coli</i> isolates from the LTEE population.....	159
Table A2. Strains constructed in this study.....	160

## List of Figures

Figure 1.1. Cellular mechanisms that contribute to genetic instability .....	12
Figure 2.1. Periodic Reselection for Evolutionarily Reliable Variants (PResERV)	
method .....	47
Figure 2.2. PResERV applied to an <i>E. coli</i> plasmid.....	48
Figure 2.3. Mutations in PResERV strains .....	50
Figure 2.4. Evolutionary stability and mutation rates in PResERV and reconstructed	
strains.....	51
Figure 2.5. Plasmid mutation rates in PResERV strains .....	53
Figure 2.6. Plasmid and chromosomal mutation rates in PResERV and reconstructed	
strains.....	54
Figure 2.7. Plasmid copy number and GFP fluorescence in PResERV and	
reconstructed strains .....	55
Figure 3.1. PResERV evolved strains have reduced mutation rates .....	84
Figure 3.2. Plasmid copy number and growth rates of PResERV strains.....	85
Figure 3.3. Mapping of causative PResERV mutations.....	87
Figure 3.4. <i>sdhA</i> and <i>sucD</i> antimutator alleles .....	88
Figure 3.5. Antimutator <i>sdhA</i> and <i>sucD</i> alleles function to relieve oxidative stress.....	89
Figure 3.6. Antimutator alleles reduce mutation rates when ported and combined .....	90
Figure 4.1. Inactivation of pSKO4 in <i>E. coli</i> .....	105
Figure 4.2. FACS sorting for MDS42 PResERV evolved populations .....	106
Figure 4.3. GFP Stability of PResERV evolved isolates .....	107
Figure 4.4. Fluorescence distributions of MDS42 controls during a GFP decay	
experiment.....	109
Figure 4.5. Fluorescence distributions of candidate antimutator strains.....	110

Figure 4.6. Mutation rates of MDS42 PResERV evolved clones .....	112
Figure A1. Evolution of rudimentary citrate utilization by activating <i>citT</i> expression is slightly beneficial in the genetic background in which it evolved and in the LTEE ancestor .....	152
Figure A2. Fitness consequences of evolving Cit <sup>+</sup> in different evolved genetic backgrounds .....	154
Figure A3. Potential for evolving Cit <sup>+</sup> mapped onto phylogeny .....	155
Figure A4. Changes in the potential for innovation along the lineage leading to Cit <sup>+</sup> due to genetic and population factors .....	157
Figure B1. DIY gellan gum plates.....	167



## **Chapter 1: Introduction**

The future exists in our ability to domesticate biology and to use it as a tool to address societal challenges. Within the last decade, this notion has been widely discussed and there have been substantial efforts, with an increasing number of successes, aimed at using biology solve problems in areas such as drug development, bioenergy, production of industrial enzymes and chemicals, and agriculture<sup>1-6</sup>. The nascent stages of this effort consisted of devising solutions to a basic question – how can we make biology easy to engineer? One initial strategy was to regard biology through the lens of established engineering disciplines, such as electrical or software engineering. If a cell could be programmed as if it were a computer, then it could be engineered by using a standard set of parts. These parts would consist of the DNA elements needed for a given function to occur (promoters, genes, terminators, ribosome-binding sites, etc.) and could be combined to create more complex synthetic circuits or pathways. To process a desired output, the DNA encoding a new function is placed into a host organism, or chassis. This is an attractive approach due to the complementary aspects of the hybrid system. Biological systems are valuable in that they are inherently able to reproduce, but they lack standardization, predictability, and reliability. Engineered systems are the opposite. They are generally straightforward to construct and operate, but they are designed to be discarded after use and lack the ability to reproduce. If this idea could be realized to its full potential, and these fields were successfully merged, one could program cells to be replicating computers<sup>1</sup>.

However, a computer is not always the best analogy for a cell. Cells are highly complex systems which are inherently ephemeral. Therefore, engineering biology is challenging. While recombinant DNA technology is widespread in research applications,

there have been very few marked successes in biological engineering that directly impact the daily lives of most people. Some of these examples are microbial synthesis of insulin, recombinant antibodies, the anti-malarial drug artemisinin, and more recently, the HPV vaccine. The first of these technologies was the commercial production of human insulin in *Escherichia coli* in 1978. Prior to this development, insulin was obtained from the pancreas glands of animals slaughtered for food, in which 8,000 pounds of animal pancreas were needed to produce one pound of insulin. Microbial production of insulin was seminal because it provided diabetic patients with human insulin instead of animal insulin, which was known to cause allergic reactions. Additionally, the dependency on large quantities of animal glands was eliminated. Production of human insulin in *E. coli* was achieved by inserting genes encoding for the two chains, “A” and “B”, and inducing their synthesis for purification. Each chain was independently purified and then co-incubated to produce an active protein by disulfide bond formation<sup>7</sup>.

The more recent achievement of the production of the anti-malarial drug artemisinin in yeast in 2014 required extensive engineering and rewiring of the host metabolic network to generate high-titers at a low cost. Artemisinin extraction is traditionally plant-based and therefore there are drastic fluctuations in the availability and price of the drug. Engineering efforts were initially galvanized by a worldwide shortage of artemisinin. Production of artemisinin occurs by synthesis of a precursor, artemisinic acid, in an industrial *Saccharomyces cerevisiae* strain, CEN. PK2, followed by purification and chemical conversion to artemisinin. This process took ~10 years to develop<sup>8</sup>. Today, yeast-based artemisinin is not yet commercially available due to price drops in plant-based artemisinin, but it is envisioned as a substitute when plant-based prices increase<sup>9</sup>.

## **METABOLIC COST AND ADAPTATIONS ASSOCIATED WITH ENGINEERING BIOLOGY**

Cells have a remarkable ability to adapt. Therefore, it is not surprising that routine genetic modifications, such as insertion of a deletion cassette into a chromosome, can cause a genomic imbalance that will give rise to compensatory effects. For example, deletion of a gene that causes a growth defect may incite the occurrence of mutations that restore the growth rate<sup>10</sup>. These secondary mutations may evolve rapidly to compensate for the change and one may not be studying the intended modification or synthetic device. This is frustrating for all aspects of scientific research – testing hypotheses, performing experiments, analyzing results, reproducing results, making conclusions, etc. Furthermore, unwanted adaptive mutations are a side effect in many genetic engineering techniques, so these unplanned changes in the DNA of an organism can happen even when there is not a large selection pressure for compensatory evolution<sup>11</sup>.

Plasmids are workhorses for recombinant DNA technologies and biological engineering, even though they are known to cause a metabolic load on the host cell<sup>12</sup>. The nature of the burden imposed by engineered plasmids has recently begun to be elucidated. Intuitively, high copy plasmids are more costly than low copy plasmids, and perhaps a low copy plasmid could be made more costly by increasing plasmid protein production via stronger promoters driving gene expression and/or encoding more genes on the plasmid. There is some evidence to suggest that the nature of the plasmid-encoded elements, also known as plasmid composition, is outweighed by the plasmid copy number. Transcriptome profiles of *E. coli* cells were assayed to determine the effect of harboring two synthetic circuits encoded on low and medium copy plasmids. Profiles of cells containing the same copy number clustered together, indicating that, for these two circuits, the plasmid composition was not as important as the copy number in terms of the host cell response. Additionally, the predominant differences in gene expression between the circuits and an

empty plasmid control were in genes involved in amino acid biosynthesis. These results suggest that the metabolic load of the synthetic circuits is at the protein production level and that the host amino acid pools are limited<sup>13</sup>.

Synthetic devices rely on host machinery for translation, which is a costly process requiring many ATP. There is some evidence to propose that when heterologous devices are inserted into cells, the main factor determining how burdensome they are is their impact on ribosome availability<sup>14-18</sup>. Modeling of the gene expression process in *E. coli* was used to determine that cells can tolerate up to an additional 46% of heterologous RNA and 33% of heterologous protein in order to maintain a normal growth rate. This model was used to predict the growth rates of cells containing plasmids differing in copy number. Many properties relating to DNA replication, transcription, and translation were considered in this model, and it was found that ribosomes were the limiting cellular resource in all scenarios<sup>14</sup>. These effects are also supported with experimental data. It was found that gratuitous overexpression (~30% of the total protein content) of either  $\beta$ -galactosidase or a nonfunctional EF-Tu caused ribosomal depletion and a decrease in protein synthesis capacity after induction<sup>17,18</sup>. Moreover, overexpression of a  $\beta$ -galactosidase was found to sequester free ribosomes away from translating other cellular proteins<sup>18</sup>. Therefore, overexpression of a protein that is not beneficial to host cell fitness can cause competition for limiting ribosomes and decrease the expression of other proteins that are likely needed for growth<sup>15</sup>.

Lower growth rates can also be due to changes in energy metabolism. Transcriptome profiles of *E. coli* DH5 $\alpha$  cells grown in a bioreactor and bearing an overexpression plasmid were shown to have reduced expression of glycolytic genes<sup>19</sup>. Furthermore, growth defects associated with protein overexpression can be corrected by overexpression of *zwf*, glucose-6-phosphate dehydrogenase, the branchpoint from

glycolysis into the pentose phosphate pathway. Increasing flux through the pentose phosphate pathway may alleviate a bottleneck in the cell's reducing power and/or ability to synthesize building blocks for growth<sup>20</sup>. Another study demonstrated that during a high density fermentation involving plasmid-based overproduction of a protein, plasmid-free cells will arise due to segregation and are better able to compete with plasmid-bearing cells because these cells are impaired in their glucose uptake ability<sup>21</sup>. Over time, protein production ceases because the plasmid-free cells outcompete the plasmid-bearing cells.

One common way in which populations of cells will adapt to these changes induced by carrying a plasmid is by mutating the source of the cost. Homologous recombination can lead to rapid inactivation of a synthetic construct. In one case, the cost of a plasmid encoding GFP was alleviated by recombination between two identical transcriptional terminators flanking the *gfp* gene<sup>22</sup>. Mobile genetic elements can also alleviate unwanted load by mutating the gene responsible for the metabolic cost. For example, non-producer cells have been found to evolve within a population of cells engineered to produce mevalonic acid. The non-producers evolved because of insertion element transposition events that interrupted the genes in the mevalonic acid production pathway<sup>23</sup>. Additionally, synthetic circuits experience escape mutants. Expression of a population control circuit, in which an *E. coli* population is controlled by quorum-sensing and a toxic protein, produces escape mutants within 3-6 days of induction<sup>24</sup>. In order to reliably engineer biology, these sorts of inactivations need to be prevented

## **CELLULAR PATHWAYS THAT CONTRIBUTE TO GENETIC INSTABILITY**

Instability stems from genome rearrangements and point mutations, that can either be silent or give rise to changes such as variation in genome size, disruption in coding sequences,

appearance of new sequences, translocations of sequences to different loci, etc. Mobile genetic elements are well-known offenders of these processes. Homologous recombination between related and repeated sequences within a single chromosome can also contribute. These events can cause gene conversion, inversions, duplications, amplifications, deletions, and translocations<sup>25</sup>. In addition, normal cellular functions such as replication, cellular metabolism, and transcription can lead to genome instability (**Figure 1.1**).

Despite the great accuracy with which DNA is replicated by some polymerases, errors are nevertheless introduced into the newly synthesized template. Mismatching can happen during replication if the error is uncorrected by the polymerase proofreading domain or the DNA mismatch repair system (MMR). Whether an error is corrected can be dictated by the local sequence context with variables such as base stacking and GC content having an effect. Additionally, mutations can occur because of strand misalignment during replication. Template switching or strand slippage at sequences that have a high propensity to form hairpins (such as quasipalindromes) and long homopolymer runs lead to indels, base pair substitutions, and structural rearrangements. However, the most frequently made error by DNA polymerases is misincorporation of ribonucleotides. This error is largely precipitated by the imbalance in cellular nucleotide pools, in which NTPs are more abundant than dNTPs. Misincorporation of ribonucleotides leads to base substitutions, and surprisingly, these errors have a mild effect on mutation rate because they are almost always repaired by nucleotide excision repair (NER)<sup>26</sup>.

Free nucleotides are highly susceptible to reactive oxygen species (ROS), especially guanine. Oxidized guanine forms 8-oxo-7,8-dihydroguanine (8-oxo-GTP) which can be incorporated into DNA by main replicative polymerases (for example, Pol III in *E. coli*) and genomic guanine can also be converted to 8-oxo-G by ROS. Sanitizing enzymes such as MutT (in *E. coli*) will eliminate the incorporation of 8-oxo-GTP by

converting it to 8-oxo-GMP, therefore acting as a defense against oxidized nucleotides<sup>27</sup>. However, if unrepaired, 8-oxo-G will pair with adenine instead of cytosine and lead to either A-T→C-G or G-C→T-A transversions<sup>26</sup>. Active oxygen species that are produced by aerobic metabolism are highly mutagenic. In *E. coli*, under normal aerobic conditions, it is estimated that ~3000-5000 oxidative DNA lesions per cell division are generated. Given the high production rate, oxidative stress may be the most prominent source of spontaneous mutations<sup>27</sup>.

Cells contain an array of mechanisms for repairing DNA. As described above, MMR functions in post-replicative correction that removes problematic nucleotides immediately, before the next round of DNA replication occurs and they lead to mutations. Therefore, many native errors made by DNA polymerase are remedied. Other types of DNA lesions are commonly repaired by base excision repair (BER) and NER. BER will replace aberrant bases such as misincorporated 8-oxo-GTP or uracil, abasic sites, or single-strand breaks by excising the damaged base via a DNA glycosylase, leaving an abasic site which is subsequently cleaved by an endonuclease, and repaired by DNA synthesis and ligation. NER repair targets bulky lesions that distort the integrity of the DNA helix, the most important of these lesions being pyrimidine dimers produced by UV. NER involves an initial recognition of the lesion by scanning DNA for helical distortions, an opening of the DNA helix near the lesion to remove the single-stranded segment containing the lesion, and the remaining gap is filled by DNA synthesis and ligation<sup>28</sup>. Many lesions produced by DNA damage will result in a block of normal genome replication. When this occurs, cells will use specialized polymerases that will perform translesion synthesis (TLS) to bypass lesions damaged by thymine dimers, bulky adducts, and abasic damage. In *E. coli*, there are three TLS polymerases encoded by *polB* (Pol II), *dinB* (Pol IV), and *umuCD* (Pol V)<sup>27</sup>.

These systems target errors that directly occur on a DNA template. However, transcription can also be a source of genetic instability<sup>29,30</sup>. There are several mechanisms involved in how transcription leads to damages of the primary DNA template. Given that replication and transcription occur on the same template, collisions can occur between the replisome and RNA polymerase and will lead to severe DNA distortions (one such conformation is known as a “chickenfoot”). In processing these distortions, double-stranded breaks can be initiated. Other distortions, such as R-loops can form during transcription. R-loop formation happens when the newly transcribed RNA strand anneals back to the DNA template, forming a RNA:DNA hybrid. This exposes the non-template DNA strand as a single strand of DNA that is more susceptible to mutagenic chemicals, strand breaks, and secondary structures. These exposed DNA strands can range up to 1 kb in length. The level of transcription is important in these mechanisms, as highly transcribed genes experience more transcription-associated mutations<sup>29,30</sup>.

Our goal is engineer host organisms with an increased genetic stability. Heterologous devices have a high propensity for inactivation by the host cell due to the imposed metabolic burden associated with device expression. Therefore, there is strong selection for inactivating mutations that will render the device non-functional. However, by engineering stable cells with a reduced basal mutation rate, we can lower the probability of inactivating mutations. As an initial step toward this effort, we need to understand previously identified mechanisms that resulted in cells with reduced mutation rates as compared to wild type, also known as antimutators.



## PREVIOUSLY DESCRIBED ANTIMUTATORS

Interest in antimutators initially began because they were regarded as a powerful method to map the origins of spontaneous mutations. Understanding the mechanism by which an antimutator reduces spontaneous mutations can provide a direct link for how mutations occur in normal cells. However, mutators are historically the system of choice for studying mutational mechanisms. For example, by mapping the mutational spectra of *E. coli* mutants defective in various repair pathways, it was determined that oxidative stress was the main contributor to spontaneous mutations. Antimutators can provide insight about the source and nature of mutations that actually occur, instead of mutations that occur in the absence of certain pathways that induce elevated mutation rates<sup>31,32</sup>. Antimutators are difficult to study. There exist few selection strategies to isolate them and as a result, have mainly been identified through tedious screening methods or as suppressors to a mutator. Additionally, many antimutators tend to have a fitness cost<sup>33</sup>.

The first antimutators to be identified were DNA polymerase (Gene 43) variants in bacteriophage T4<sup>34</sup>. Biochemical analysis of these mutants determined that the antimutator phenotype was due to a tradeoff causing increased proofreading efficiency with a subsequent decrease in polymerization rate. This effect has a deleterious effect on fitness<sup>31,33</sup>. Other antimutator polymerase variants have been identified in *E. coli* DNA Polymerase III and DNA Polymerase I<sup>35-37</sup>. For Pol III, the main replicative polymerase in *E. coli*, antimutator variants were found in the  $\alpha$  subunit, *dnaE*, which catalyzes polymerization<sup>34,38</sup>. These variants are suggested to possess improved replication fidelity by increased nucleotide selectivity or better dissociation from terminal mismatches, allowing access to repair machinery for rapid correction of the error<sup>31,33</sup>. Unlike Pol III, Pol I is not essential. It participates in lagging strand synthesis, DNA repair, and replication of ColE1 plasmids. Antimutator variants of Pol I are purported to have enhanced nucleotide

selectivity or decreased extension of termini in the case of a mispaired primer<sup>33</sup>. In sum, DNA polymerases can be mutated, resulting in an antimutator phenotype.

Alternatively, removal of error-inducing systems will also lead to reduced mutation rates. A reduced-genome strain of *E. coli*, MDS42, was improved by deleting the three DNA polymerases induced during the SOS response (Pol II, Pol IV, and Pol V). These polymerases are known to induce mutagenesis and therefore, deletion of these genes reduced the mutation rate of MDS42 even further. Removal of error-prone repair will promote the use of error-free pathways for repairing damaged lesions<sup>39</sup>. Repair and housekeeping protein variants have also been linked to an antimutator phenotype. MMR activity is decreased in stationary phase *E. coli* cells, indicating that these cells likely suffer a high propensity for DNA damage. Overproduction of the MMR protein, MutL, was shown to decrease mutation rates of *E. coli* cells in stationary phase. Therefore, it is possible that other repair proteins may lead to a similar phenotype<sup>40</sup>. Regarding housekeeping proteins, overexpression of *nudG*, a Nudix hydrolase with high specificity for oxidized pyrimidines, decreased mutation frequencies under normal growth conditions and in the presence of oxidative stress, induced by hydrogen peroxide. Thus, overexpression of *nudG* better sanitizes the nucleotide pool and therefore prevents incorporation of damaged nucleotides into DNA<sup>41</sup>.

This dissertation describes mutations in other genes that lead to antimutator mechanisms identified via a directed evolution approach, Periodic Reselection for Evolutionarily Reliable Variants (PResERV). This methodology allows us to probe mechanisms that are non-obvious targets for reducing mutation rates, including mutations that do not have any trade-offs that would make cells containing them undesirable in biotechnology applications. PResERV enabled for identification of antimutator alleles in various pathways – DNA replication (**Chapter 2**), RNA processing (**Chapter 2**), and

central carbon metabolism (**Chapter 3**). Additionally, PResERV was performed on an already improved strain, MDS42, which resulted in further improvements and several interesting candidate mutations for follow-up studies (**Chapter 4**). This work expands our knowledge of genetic stability in *E coli* under typical laboratory conditions. Additionally, these antimutator strains can be used as hosts for biological engineering, which will inhibit the inactivation of engineered devices and ultimately allow engineers to tackle challenging problems.



## **Chapter 2: Directed evolution of *Escherichia coli* with lower-than-natural plasmid mutation rates**

**This chapter is reproduced (with minor modifications) from its initial publication:**

Deatherage DE, **Leon D**, Rodriguez AE, Omar S, Barrick JE. (2018) Directed evolution of *Escherichia coli* with lower-than-natural plasmid mutation rates. *Nucleic Acids Res* (in press).

### **Research Contributions**

Deatherage DE, Leon D, Rodriguez AE, and Omar S performed experiments. Deatherage DE, Leon D, and Barrick JE analyzed data. Deatherage DE, Leon D, and Barrick JE wrote the majority of the manuscript.

### **Acknowledgements**

We thank Sean Sleight and Herbert Sauro for the gift of pSKO4 and helpful discussions; Drew Tack and Andy Ellington for providing pTEM-1.D254tag; Stratton Georgoulis for contributing to the development of the numerical simulations; and four anonymous reviewers for helpful comments. We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources.

## ABSTRACT

Unwanted evolution of designed DNA sequences limits metabolic and genome engineering efforts. Engineered functions that are burdensome to host cells and slow their replication are rapidly inactivated by mutations, and unplanned mutations with unpredictable effects often accumulate alongside designed changes in large-scale genome editing projects. We developed a directed evolution strategy, Periodic Reselection for Evolutionarily Reliable Variants (PResERV), to discover mutations that prolong the function of a burdensome DNA sequence in an engineered organism. Here, we used PResERV to isolate *E. coli* cells that replicate ColE1-type plasmids with higher fidelity. We found mutations in DNA polymerase I and in RNase E that reduce plasmid mutation rates by 6- to 30-fold. The PResERV method implicitly selects to maintain the growth rate of host cells, and high plasmid copy numbers and gene expression levels are maintained in some of the evolved *E. coli* strains, indicating that it is possible to improve the genetic stability of cellular chassis without encountering trade-offs in other desirable performance characteristics. Utilizing these new antimutator *E. coli* and applying PResERV to other organisms in the future promises to prevent evolutionary failures and unpredictability to provide a more stable genetic foundation for synthetic biology.

## INTRODUCTION

Populations of cells engineered to function as factories or biosensors experience a failure mode that is peculiar to living systems: they evolve. Unwanted evolution is a foundational problem for bioengineering that limits the efficiency and predictability of metabolic and genome engineering efforts<sup>10,42-45</sup>. Often an engineered function diverts critical resources from cellular replication or otherwise interferes with growth or homeostasis<sup>11,46</sup>. In these cases, ‘broken’ cells with mutations that inactivate the engineered function can rapidly outcompete the original design<sup>5,16,47</sup>. The rate at which an engineered function decays within a cell population in this manner can be summarized as an evolutionary lifetime or half-life<sup>5</sup> or defined in terms of an evolutionary landscape by the rates at which various mutational failure modes occur and their respective fitness benefits<sup>16,47</sup>.

It is sometimes possible to edit a genome to eliminate or reduce the rate at which certain types of mutations occur<sup>23,39,48-51</sup> or to devise a way of reducing the burden of an engineered function<sup>42,52</sup>. However, given the complexity of DNA replication and repair processes and the multifarious ways that an engineered function can burden a host cell, a point is generally reached at which it is difficult to further improve upon the reliability of a cell. Directed evolution is an effective strategy for optimizing the performance of complex systems with many interacting components, even when they include unknown factors. For example, it has been used to engineer novel enzymes that outperform their natural counterparts<sup>53</sup> and to tune artificial gene circuits to effectively perform logic operations<sup>54</sup>.

Given the similarly complex constraints underlying cellular mutagenesis and the fitness burdens of diverse engineered functions, we reasoned that a directed evolution procedure, Periodic Reselection for Evolutionarily Reliable Variants (PResERV) (**Fig. 2.1**), could be an effective strategy for improving the evolutionary reliability of an

engineered cell. In PResERV, one artificially selects for mutant cells that exhibit improved maintenance of a burdensome engineered function over tens to hundreds of cell divisions. We expected that PResERV might isolate cells with mutations that either reduced the rate at which failure mutations occurred or the fitness burden of the engineered function, or both, possibly in ways that would generalize to stabilizing other engineered functions in the evolved cells. Here, we describe *E. coli* strains evolved by PResERV that exhibit lower-than-natural mutation rates for genes encoded on high-copy plasmids, thereby stabilizing them against unwanted evolution.

## RESULTS

### **PResERV experiment with a ColE1 plasmid in *E. coli***

We applied PResERV to *E. coli* K-12 strain BW25113<sup>55</sup> transformed with pSKO4, a high-copy-number pBR322 plasmid (ColE1-type origin) that encodes GFP under control of an inducible promoter<sup>22</sup>. GFP expression is a generic proxy for a costly engineered function in this scenario. A UV-mutagenized library of cells containing pSKO4 was propagated through daily 1:1000 serial transfers in the presence of antibiotic selection for plasmid retention. Under these conditions, cells with mutations in pSKO4 that inactivate or reduce costly GFP expression evolve, outcompete fully fluorescent cells, and constitute a majority of the population within a few days<sup>22</sup>. GFP fluorescence of cells in the PResERV population was periodically monitored by flow cytometry using the red fluorescent nucleic acid dye SYTO17 as a counterstain to improve detection of cells with a low GFP signal. When 75% or more of the cells exhibited reduced GFP signal, cell sorting was used to isolate  $\sim 10^5$  cells that remained at least as fluorescent as the ancestor to continue the



population. We subjected this population to a total of 8 sorts spread throughout 30 regrowth cycles (**Fig. 2.2A**).

Six *E. coli* clones designated AER7–AER12 were isolated from the final population for further characterization. Five of these maintained more fully-fluorescent cells for more cell doublings than the unevolved wild-type strain with the wild-type pSKO4 plasmid (**Fig. 2.2B**). Mutations in the plasmid, the *E. coli* chromosome, or both could have been responsible for these improvements. To determine which was the case, we cured these cells of their plasmids and retransformed them with the wild-type pSKO4 plasmid, and we also isolated plasmids from each of the evolved strains and transformed them into unevolved wild-type *E. coli* cells. For four of these strains (AER7, AER8, AER9, and AER12), the improvement in the evolutionary lifetime of GFP expression appeared to be mainly due to mutations in the *E. coli* chromosome rather than mutations in the pSKO4 plasmid (**Fig. 2.2C**).

### **Mutations in PResERV strains**

We sequenced the genomes of these four evolved clones to understand the genetic basis of their improved reliability (**Fig. 2.3**). In agreement with the re-transformation tests, no mutations were found in the pSKO4 plasmid in any of these strains. Each contained from four to ten mutations in the *E. coli* chromosome. These mutations could theoretically lead to the improved maintenance of GFP expression that we observed by reducing the burden of GFP expression from the plasmid or by reducing the rate at which mutations that inactivate GFP arise. Therefore, we examined the lists of mutations in these strains to see if they hit any genes known to be involved in these processes.

Two of these strains (AER7 and AER8) had eight identical mutations while a third strain (AER9) had these same eight mutations plus two additional ones. All three shared mutations in *polB* and *rne* that were candidates for affecting evolutionary stability. PolB (Pol II) is a stress-induced DNA polymerase that participates in translesion synthesis and nucleotide excision repair. The *polB* gene sustained two mutations in these PResERV strains: a missense mutation (H597Y) and a nonsense mutation earlier in the reading frame (S558\*). The full-length PolB protein is 783 amino acids in length, and the stop codon mutation truncates the protein within its catalytic core<sup>56</sup>. Presumably, this results in a complete loss of Pol II activity in the mutant. Deletion of *polB* in the clean-genome *E. coli* strain MDS42 has been shown to lead to ~30% lower chromosomal mutation rates<sup>39</sup>. The *rne* gene (encoding RNase E) contains a missense mutation (L222S) in all three strains. RNase E regulates the copy number of ColE1 origin plasmids in *E. coli* by processing the RNA I antisense regulator of the RNA II replication primer<sup>57,58</sup>. Cells defective in *rne* accumulate higher levels of RNA I and have reduced plasmid copy number<sup>59</sup>. The site of the PResERV mutation is within the RNaseH-like domain of RNase E, which is involved in determining its RNA substrate selectivity, but its effect on the activity of this enzyme is not clear from the structural context<sup>60</sup>.

The fourth sequenced strain (AER12) had a completely different set of four mutations, which included a missense mutation in *polA* (H734Y). PolA (Pol I) is the DNA polymerase that is utilized primarily for filling gaps during lagging strand synthesis and in DNA repair in *E. coli*. It is also responsible for extending the primer derived from RNA II during replication of ColE1-type plasmids<sup>58</sup>. Antimutator variants of PolA that lower the frequencies of mutations observed on a reporter plasmid have been identified previously by screening a sequence library created by mutagenizing an *exo<sup>-</sup>* PolA variant lacking 3'→5' exonuclease proofreading activity<sup>36</sup>. The exact same substitution that we observed

(H734Y) was found among the 592 active polymerase variants characterized in that study, but the effects of this specific mutation on polymerase function were not reported. H734 is located near the phosphate groups of the dNTP substrate when it is bound to the Klenow fragment of Pol I<sup>61</sup>, indicating that the PResERV mutation may have an effect on nucleotide binding.

### **Evolutionary stability and mutation rates in PResERV and reconstructed strains**

To test whether these three mutations contributed to the increased evolutionary reliability of the PResERV strains, we tested *E. coli* strains in which we reverted the evolved alleles back to their wild-type sequences. We then propagated replicate populations of wild-type *E. coli*, two focal evolved clones (AER12, the strain with the *polA* mutation; and AER8, one of the three strains containing mutations in *polB* and *rne*), and four revertant strains (one for each mutation and also a strain in which *polB* and *rne* were both reverted) under the same conditions as the initial evolution experiment and monitored the loss of fluorescence over the course of ~100 cell doublings (**Fig. 2.4**). Mutations in *polA* and *rne* appeared to be responsible for most or all of the improved stability, as reverting these mutations reduced the evolutionary lifetime of GFP expression back to a level similar to that observed in the wild-type strain. In contrast, reverting the *polB* mutation alone or reverting it in a strain that also had the *rne* mutation reverted did not appreciably affect how rapidly GFP expression decayed.

We next used Luria-Delbrück fluctuation tests<sup>62</sup> to determine if the increase in evolutionary reliability in these strains was associated with a decrease in mutation rates. We first measured the rates of point mutations that reverted a stop codon in a  $\beta$ -lactamase gene cloned into another pBR322-based plasmid designated pTEM-1.D254tag<sup>63</sup>. Mutation

rates to carbenicillin resistance, which requires mutating this stop codon to a sense codon, were significantly lower in each of the two focal evolved clones compared to wild type in multiple experiments (**Fig. 2.5**). In agreement with the changes in the evolutionary stability of GFP expression, reversion of either the *polA* or *rne* mutation raised the mutation rate to that of the wild-type *E. coli* strain, and reversion of the *polB* mutation had no detectable effect on the mutation rate (**Fig. 2.6A**). We also measured mutation rates in two further sets of fluctuation tests, selecting either for resistance to rifampicin or to d-cycloserine, which require mutations in genes located on the *E. coli* chromosome in both cases. We did not find any significant improvements versus the wild-type strain in these assays (**Fig. 2.6B**). Thus, it appears that PResERV discovered *E. coli* mutants that primarily display lower plasmid mutation rates, with much less of an effect, if any, on mutation rates in the chromosome.

### **Plasmid copy number and GFP fluorescence in evolved strains**

Given previous reports of lower plasmid copy number when *rne* function is reduced in a temperature-sensitive mutant<sup>59</sup> and that *polA* antimutator mutations can lead to slower rates of DNA replication<sup>36</sup>, we were concerned that a decrease in plasmid copy number in the PResERV evolved cells could give a false signal of improvement in our two assays. First, having fewer plasmids per cell would lower the GFP expression burden and thereby increase the number of cell doublings it would take for new cells that arise with mutated plasmids to outcompete cells with wild-type plasmids (i.e., it would increase the apparent evolutionary stability). Second, with fewer plasmids per cell there would be a smaller chance that any given cell would experience a mutation in one of its plasmids that would lead to resistance in the  $\beta$ -lactamase stop codon reversion assay (i.e., it would reduce the

apparent mutation rate per cell). Therefore, we measured the copy number of the pTEM-1.D254tag plasmid in the two focal evolved clones and four reconstructed strains using qPCR (**Fig. 2.7A**), and we also examined the per-cell GFP fluorescence from the pSKO4 plasmid in each strain (**Fig. 2.7B**).

We found that the *polA* mutation did reduce plasmid copy number somewhat. The evolved *polA* strain (AER12) had marginally fewer plasmids per chromosomal DNA copy when compared to wild-type ( $p = 0.0666$ , one-tailed  $t$ -test on log-transformed values) and also exhibited reduced GFP fluorescent intensity ( $p = 0.0142$ , one-tailed  $t$ -test on log-transformed GFP+ subpopulation medians). Interestingly, other mutations in the evolved strain appeared to counteract the effects of the *polA* mutation, as reverting just this mutation to the wild-type allele increased both copy number ( $p = 0.0011$ ) and GFP intensity ( $p = 0.0003$ ). GFP signal ( $p = 0.0108$ ) and perhaps copy number ( $p = 0.0893$ ) were even greater in this *polA* revertant that still contained all other evolved mutations than they were in the original wild-type *E. coli* strain. Overall, these results suggest that there is a trade-off in the evolved *polA* mutant between plasmid copy number and mutation rate.

In contrast, plasmid copy number did not vary when comparing wild-type, the evolved strain with *rne* and *polB* mutations (AER8), and all three reconstructed strains reverting those mutations singly and in combination (one-way ANOVA,  $F_{4,10} = 0.439$ ,  $p = 0.778$ ). Here, too, there was evidence that other mutations in this evolved strain may have increased GFP intensity, as all four of the AER8-derived strains considered together had indistinguishable fluorescence intensities (one-way ANOVA,  $F_{3,32} = 1.153$ ,  $p = 0.343$ ) that were, as a group, significantly greater than that of the wild-type ( $p = 0.0112$ , one-tailed  $t$ -test). Therefore, the evolved *rne* allele reduced plasmid mutation rates with no detectable trade-off in terms of plasmid copy number or gene expression.

To determine whether reduced plasmid copy number in the AER12 strain containing the *polA* mutation could explain the reduction of 20- to 60-fold in the plasmid mutation rate measured for this strain (**Fig. 2.5** and **Fig. 2.6A**), we performed numerical simulations of the growth of cell populations that included multicopy plasmid replication, mutation, and segregation (see **Methods**). Our qPCR results indicate that the copy number of the pTEM-1.D254tag plasmid was reduced from ~410 plasmids per *E. coli* chromosome in the wild-type strain to ~185 in AER12. We simulated the results of fluctuation tests with 410 mutational reporter plasmids per cell and with other parameters chosen to match the observed numbers of mutant cells per culture for each of our four different sets of mutation rate measurements comparing wild type and AER12. Then, we performed a new set of simulations with the same parameters but reducing the plasmid copy number to 185 to determine by how much this would reduce the apparent plasmid mutation rate inferred from the Luria-Delbrück analysis. We found that a reduction in copy number of 2.2-fold is expected to yield a roughly proportional change in the apparent mutation rate. The result varies slightly if one changes how many mutant plasmids in a cell are necessary for it to give rise to a mutant colony, a parameter that is unknown in our system but is likely one or a just a few plasmids with restored  $\beta$ -lactamase copies per cell. For simulations requiring one mutant plasmid per cell we predicted a reduction of 2.00-fold (1.93–2.07, 95% bootstrap confidence interval, see **Methods**) in the apparent mutation rate in AER12. For three copies the reduction was 2.12-fold (2.05–2.20), and for ten copies it was 2.37-fold (2.24–2.51). We conclude that the reduction in plasmid copy number in AER12 is not sufficient to explain a majority of the reduction in plasmid mutation rates in that strain.

## DISCUSSION

Mutation rates in microbial populations reflect a dynamic balance between different evolutionary forces and inherent constraints on organisms that have DNA as their genetic material. On one side, there is a universal selection pressure to minimize mutation rates because most new mutations are far more likely to be deleterious to fitness than beneficial<sup>64-66</sup>. This risk associated with deleterious mutations contributes to genetic load. That is, there is a fitness cost associated with a given mutation rate in terms of the fraction of an organism's offspring that will experience lethal or deleterious mutations that lead to their immediate or eventual extinction. If selection to reduce genetic load were the only evolutionary force in play, then a mutation rate of zero would be optimal. On the other side of the balance, there are at least three different forces or barriers that will prevent the evolution of lower mutation rates past a certain point in microbial populations: second-order selection for evolvability and limits imposed by the strength of genetic drift and physiological constraints.

New mutations may be a bad bet on average, but they do—more rarely—generate beneficial genetic diversity that is necessary to fuel adaptive evolution. Thus, under certain circumstances, mutation rates can evolve to rebalance the potential for beneficial mutations against the risk of deleterious mutations<sup>67-69</sup>. For example, laboratory populations of bacteria and yeast often evolve hypermutation (elevated mutation rates)<sup>70,71</sup> because they experience strong and constant selection pressures that can indirectly favor more evolvable lineages that have a greater chance of sampling rare adaptive mutations<sup>72,73</sup>. The simplicity of laboratory environments compared to nature also means that there is less of a deleterious genetic load associated with evolving a high mutation rate in these experiments. Many mutations that would be lethal under other circumstances (e.g., that disrupt pathways for utilizing alternative nutrients or stress responses for contingencies that are never

experienced) will be effectively neutral in the comparatively monotonous environments of these experiments<sup>74,75</sup>. For similar reasons, hypermutators also commonly evolve in populations of bacteria during the long-term progression of chronic infections treated with antibiotics, such as for *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients<sup>76-78</sup>.

The molecular basis for the evolution of bacterial hypermutators in the laboratory and in the clinic is usually straightforward. Mutations disrupt major housecleaning enzymes (e.g., *mutT*) or DNA repair pathways (e.g., *mutS*), often leading to an increase in point mutations with a characteristic base substitution spectrum<sup>70,71,77,79</sup>. Interestingly, some experiments have shown that experimental populations that are started with or that spontaneously evolved hypermutation can subsequently evolve reduced mutation rates<sup>80-83</sup>. This can occur when hypermutator populations are propagated through severe population bottlenecks, which exacerbates the genetic load associated with a given mutation rate while reducing the chances of sampling adaptive mutations<sup>80,81</sup>. Reduced mutation rates have also been observed to evolve after a population becomes well-adapted to its environment and opportunities for beneficial mutations diminish relative to the risk of deleterious mutations<sup>82,83</sup>. When the molecular mechanisms have been examined in detail, the evolution of lower mutation rates in these experiments has been found to occur through new mutations that partially compensate in some way for the defect in the hypermutator<sup>82</sup>, or by exact reversion of the mutation responsible for hypermutation<sup>83</sup>. The evolution of cells with a mutation rate that is lower than that of the ancestral, wild-type microbe has not been observed in these experiments.

Despite the potential for evolving hypermutation, wild-type microbes isolated from nature almost always have very low mutation rates<sup>85</sup>. The uniformity of these baseline rates across many species isolated from diverse environments suggests that a different balance



of evolutionary forces than the one between genetic load and the potential for beneficial mutations is normally responsible for setting mutation rates in nature. If one assumes that wild-type microbes are already well-adapted to the combinations of complex and varying environments that they regularly experience, then there may be little or no benefit possible from further mutations. Under these circumstances there will only be selection to minimize genetic load. What then would set the lower bound on mutation rates? Because baseline mutation rates have been found to scale inversely with effective population sizes across many organisms, it has been argued that genetic drift is the dominant force opposing selection for even lower mutation rates<sup>84,85</sup>. This ‘drift barrier’ arises because once the mutation rate is sufficiently low, the very small and indirect marginal benefit for a new mutation that leads to an even lower mutation rate becomes so insignificant that it looks effectively neutral to natural selection. That is, natural selection does not have the power to favor this hypothetical new antimutator allele such that it will reliably increase in frequency on its merits and eventually fix in the population.

Another potential barrier to the evolution of lower mutation rates considers the molecular biology of DNA replication and repair. Biochemical and genetic studies of bacteria over the past several decades have mapped a complex suite of pathways dedicated to maintaining genome integrity via overlapping and redundant mechanisms<sup>86</sup>. There is a direct fitness cost to a cell for expressing any protein<sup>11,46</sup>, and there may be other fitness costs associated with increased surveillance for DNA damage and enforcing replication fidelity, such as off-target promiscuous activities of housecleaning enzymes<sup>87</sup> or slower rates of DNA polymerization in exchange for increased proofreading<sup>36</sup>. Thus, there must reach a point at which the direct fitness cost of evolving additional molecular machinery outweighs the diminishing indirect benefit of further reducing the genetic load from deleterious mutations. However, there is no evidence that this physiological lower limit on

mutation rates has been reached in natural microbes. Antimutators with reduced point mutation rates have been identified by genome-wide genetic screens<sup>37</sup>, targeted mutagenesis<sup>35,36</sup>, gene disruption<sup>39</sup>, and gene overexpression<sup>41</sup>. Some of these antimutators do exhibit growth trade-offs<sup>36,37</sup>, but some do not appear to have any deleterious side-effects<sup>36,39,88</sup>, at least in the laboratory environments in which they have been tested. Yet, due to the intrinsic instability of DNA, which can be chemically damaged or miscopied in an dizzying variety of ways<sup>86</sup>, there must exist some finite, non-zero mutation rate at which this physiological genetic stability limit is reached.

In this study, we show that imposing artificially strong selection for bacterial cells that are less likely to give rise to mutations in a reporter gene on a plasmid can overcome the selection pressures and other barriers that normally oppose the evolution of reduced mutation rates. Specifically, we developed and used a Periodic Reselection for Evolutionarily Reliable Variants (PResERV) directed evolution approach to isolate *E. coli* host strains with mutations in their chromosomes that lead to lower-than-natural mutation rates in genes encoded on high-copy vectors such as pUC and pBR322 from the ColE1 plasmid incompatibility group. We sequenced the genomes of four improved PResERV strains to better understand the molecular basis for the improvements and found mutations in three key genes (*polA*, *polB*, and *rne*). Then, we characterized the effects these mutations had on the evolutionary stability of burdensome GFP expression, on plasmid and chromosomal mutation rates, and on plasmid copy number and gene expression in the evolved *E. coli* strains.

One PResERV strain had a mutation in DNA polymerase I (*polA*) that reduced plasmid mutation rates by ~30-fold. Pol I is required for the normal replication of ColE1 plasmids in *E. coli*<sup>58</sup>. Both hypermutator and antimutator variants of this polymerase have been shown to affect the fidelity of DNA replication<sup>36,89,90</sup>, so it was not surprising that we

identified a mutation in *polA* that increased genetic stability. In fact, the exact amino acid substitution in Pol I recovered by PResERV (H734Y) was found previously in a library of mutagenized  $\text{exo}^-$  Pol I sequences<sup>36</sup>. Though the specific effects of the H734Y mutation were not reported individually in this study, many of these mutagenized Pol I variants were antimutators. Their improved fidelity was attributed to increased selectivity for the incoming nucleotide. The *polA* mutation that we recovered may act similarly, as it is located close to the binding site for the incoming dNTP<sup>61</sup>. Thus, our identification of a *polA* mutation was in essence a positive control that PResERV could successfully isolate generalizable antimutator alleles, as opposed to mutations that increased the evolutionary stability of just the pSKO4 plasmid in an idiosyncratic way (e.g., by reducing the cost of expressing GFP).

Pol I has an outsized role in replicating ColE1 plasmids compared to its relatively minor roles in lagging-strand synthesis and DNA repair during normal replication of the *E. coli* chromosome. Pol I initiates plasmid DNA replication by extending a primer that is processed from the RNA II transcript derived from the plasmid origin. In the canonical model of plasmid replication, the DNA polymerase holoenzyme involved in chromosomal replication, which utilizes Pol III, takes over plasmid replication after Pol I extends the primer by ~400-500 nucleotides<sup>58</sup>. However, there is evidence that Pol I also replicates other portions of ColE1 plasmids, at least some of the time. When a hypermutator Pol I variant was expressed in *E. coli* cells, plasmid mutation rates were most elevated close to the origin of replication, within the expected 400-500 base-pair window, but mutation rates were still much higher than normal within a region extending at least 3700 base pairs downstream of the origin<sup>90</sup>.

The reading frame for GFP on the pSKO4 plasmid used in PResERV is located from ~350 to ~750 base pairs downstream of the origin (as measured from the typical end

of the RNA II transcript after nucleolytic processing). This places at least part of the GFP gene within the region known to be heavily replicated by Pol I, meaning that we would expect to observe a particularly strong effect of a *polA* antimutator allele during PResERV and in subsequent decay experiments in which we monitored the evolutionary stability of GFP expression over multiple growth cycles. In contrast, the stop codon in  $\beta$ -lactamase on the pTEM-1.D254tag reporter plasmid that we used in fluctuation assays to measure plasmid mutation rates is located ~3000 base pairs downstream of the end of RNA II. We still see greatly reduced mutation rates in this reporter, corroborating the prior observations that changes in Pol I fidelity impact mutation rates across most or all of the sequence of a ColE1-type plasmid. Also in broad agreement with previous studies, which report that Pol I variants have less of an effect on chromosomal mutation rates compared to plasmid mutation rates<sup>89,90</sup>, we found no significant difference in chromosomal mutation rates from the PResERV *polA* antimutator allele.

Our overall goal is to construct an *E. coli* cell that is more robust against unplanned evolution to serve as an improved ‘chassis’ for synthetic biology and biotechnology applications. For this purpose, the most useful antimutator alleles are those that increase the genetic stability of an engineered DNA sequence in a host cell without any trade-offs in other desirable traits. In one important respect, the PResERV *polA* mutation and many other *polA* antimutator alleles pass this test: they do not negatively affect *E. coli* growth rates<sup>89</sup>. However, the PResERV allele does exhibit a significant trade-off that diminishes its potential utility. Copy number of the pTEM-1.D254tag plasmid used to measure mutation rates was reduced by ~55% in strains with the evolved *polA* allele. This decrease is consistent with a reduction in GFP expression from the pSKO4 plasmid when the evolved *polA* mutation was present in a strain. We used numerical simulations to show that this slight reduction in plasmid copy number can explain only ~2-fold of the ~30-fold

reduction in plasmid mutation rates we observed in the evolved strain. Still, as ColE1 plasmids are widely used for cloning and protein overexpression, where maximal yield of plasmid DNA or a protein encoded on the plasmid is the primary goal, this trade-off of much lower plasmid mutation rates at the expense of reduced plasmid copy number would not be favorable, on balance, for many biotechnology applications.

Other mutations in Pol I have previously been found to reduce the copy number of high-copy ColE1 plasmid variants<sup>91</sup>. It has been hypothesized that they might have this effect by decreasing the frequency of initiation of DNA synthesis from the RNA II primer, by reducing the speed of DNA polymerization, or by some combination of the two. In biochemical assays, some of the *exo<sup>-</sup>* Pol I antimutator variants have been reported to have significantly reduced rates of polymerization<sup>36</sup>. They would presumably also reduce ColE1 plasmid copy number, though this has not been tested. However, other antimutator Pol I variants apparently retain wild-type enzyme activity<sup>36</sup>. It is possible that these *polA* mutations do not exhibit any trade-off in plasmid copy number and would ultimately be more useful than the PResERV allele for constructing an improved *E. coli* host strain.

The three other sequenced PResERV strains all shared mutations in two genes that also have known roles in DNA replication fidelity and regulation of plasmid replication: DNA polymerase II (*polB*) and RNase E (*rne*). Pol II is a repair polymerase induced by the SOS and RpoS responses<sup>92,93</sup>. We observed a mutation that creates a premature stop codon in *polB* at amino acid 558. This truncation likely results in a completely inactivated enzyme. There is also a second point mutation later in the *polB* reading frame in these strains that would result in an amino acid substitution (H597Y) if the nonsense mutation were not present. The occurrence of two nearby mutations in the same gene is probably due to our use of UV mutagenesis to create initial genetic diversity in the *E. coli* population at the beginning of PResERV, as clustered mutations can result from long-patch repair of

UV damage<sup>94</sup>. Despite the fact that UV damage induces the SOS response, we do not expect that these mutations in *polB* were favored due to any direct connection to the mutagenesis procedure. Loss of Pol II function does not appreciably affect cell survival or the overall level of mutagenesis after UV exposure unless other DNA repair pathways are also inactivated<sup>95</sup>.

The connection between Pol II activity and *E. coli* mutation rates has multiple facets. On one hand, Pol II can act as a high-fidelity alternative to the other stress-induced polymerases (Pol IV and Pol V). As a consequence, inactivation of *polB* increases the incidence of point mutations arising from the repair of DNA double-strand breaks that occur during long-term carbon starvation<sup>96,97</sup>. However, Pol II also participates in mutagenic translesion synthesis pathways that repair other types of DNA damage in an error-prone manner. Accordingly, deletion of Pol II has been reported to have the opposite effect and reduce mutagenesis associated with certain DNA base adducts<sup>98,99</sup> and in cells exposed to antibiotics that can cause DNA damage<sup>100</sup>. The mutagenic effect of *polA* activity appears to dominate in *E. coli* cells growing under standard laboratory culture conditions, as incorporating a deletion of *polB* into an engineered reduced-mutation variant of the MDS42 clean-genome *E. coli* strain lowered chromosomal mutation rates by ~30%<sup>39</sup>.

Despite these connections between Pol II and mutagenesis, the mutant *polB* gene sequence from PResERV was not associated with a significant change in mutation rates in our assays when we reverted it to the wild-type sequence in the evolved strain or in a strain in which the *rne* mutation found in the same evolved strains was also reverted. It is possible that this is due to epistatic interactions with the other five mutations common to this set of evolved strains, although none of these mutations affect genes with an obvious connection to DNA replication or repair processes. One or both of the Pol II mutations could have been

present immediately after UV mutagenesis at the beginning of PResERV and reduced mutation rates in this context. Then, subsequent mutations that arose in this winning lineage during the regrowth cycles of PResERV might have overshadowed the effect of the Pol II mutations by making them redundant. However, we believe this is unlikely to be the only explanation, as the evolved *rne* allele on its own seems to explain all of the reduction in mutation rates, whether or not the evolved *polB* sequence is present. It is also possible that some aspect of the environment experienced by cells during PResERV but not during the mutation rate assays introduced a stress that favored *polB* inactivation. For example, the PResERV cultures were often interrupted by diluting them into water and processing them through a cell sorter before the next growth cycle. In any case, it is clear that knockout of Pol II is not as effective at reducing mutation rates under normal growth conditions as the mutation in RNase E that is present in the same strains.

RNase E is an endoribonuclease with global roles in RNA maturation, processing, and decay<sup>101</sup>. It is involved in tRNA, rRNA, and small RNA processing and has been reported to initiate the decay of ~60% of *E. coli* mRNAs<sup>102,103</sup>. RNase E is also specifically involved in controlling the copy number of ColE1 plasmids<sup>57,58</sup>. It does so by cleaving the regulatory antisense RNA I transcript at a specific site, which converts it into an inactive form that cannot bind to and inhibit processing of RNA II into a productive primer. RNase E is an essential gene, but eliminating its expression using a temperature-sensitive mutant has been shown to reduce plasmid copy number, as is expected from the resulting increase in levels of the active form of the RNA I inhibitor<sup>59</sup>. Despite this connection to the regulation of initiation of ColE1 plasmid replication and unlike the *polA* allele in the other PResERV strain, the evolved *rne* allele did not significantly change plasmid copy number, as measured using qPCR for the pTEM-1.D254tag mutational reporter plasmid or in terms of the fluorescence output from the pSKO4 plasmid. This *rne* allele demonstrates an

advantage of using the PResERV directed evolution approach. Although RNase E has a known role in ColE1 plasmid replication, it would not have been an obvious target for rationally engineering a more genetically stable host strain.

The PResERV *rne* allele was responsible for a 6-fold reduction in plasmid mutation rates with no significant effect on chromosomal mutation rates. The altered amino acid (L222S) is located within its split RNaseH-like domain near the embedded 5' sensor domain that is responsible for its preference for RNA substrates with a 5' monophosphate<sup>60</sup>. It is unclear how this mutation might affect RNase E activity and lead to a reduction in plasmid mutation rates. It could potentially have a direct effect on processing of RNA I and/or RNA II that alters the balance of different DNA polymerases used to replicate and/or repair ColE1-type plasmids, though it is hard to imagine how this could happen without also affecting plasmid copy number. Alternatively, the *rne* mutation may have an indirect effect by altering the decay or maturation of other RNAs in a cell. RNase E has been shown to affect the biogenesis and activity of small RNA<sup>103</sup>, many of which are involved in stress responses<sup>104</sup>, to point out one such possibility among many. It will take future work examining the biochemical effects of this mutation on enzyme activity and its global effects on the *E. coli* transcriptome to decipher why it has an antimutator effect on plasmid replication. Of the three mutations in the PResERV strains that we studied, this mutation in RNase E appears to hold the greatest promise for applications in biotechnology, as the antimutator effect is not associated with any unwanted trade-offs in terms of growth rate or plasmid copy number in the standard culture conditions we used.

Overall, we expect that the PResERV approach will be widely applicable and useful for isolating mutations that make engineered cells more robust against evolutionary failure by lowering mutation rates. One advantage of PResERV is that it is agnostic to the source of mutations and the type of host cell. It will select for mutants that eliminate the dominant



cause of mutations inactivating the reporter gene used for cell sorting, and it can be employed iteratively to further eliminate the next-most dominant source of mutations by subsequently introducing new genetic variation into the population and continuing the cycles of cell growth and sorting. When there is genome-wide genetic variation in the cell population, PResERV can discover mutations in genes of unknown function or pathways that do not have obvious connections to mutagenesis, like the *rne* mutation in this study. In the future, PResERV could also be used on libraries of cells that target variation to one key enzyme (e.g., *polA*) or to a suite of genes known to be involved in DNA replication and repair, by using multiplex genome editing methods<sup>105,106</sup>. The current study demonstrates proof-of-principle for the PResERV approach, but by examining just six isolates from one mutagenized population, it has clearly not identified all of the ways that mutations in the *E. coli* genome can lower mutation rates.

An advantage of using directed evolution compared to screening approaches that have been used to isolate antimutators in the past<sup>37</sup> is that the cycles of regrowth between cell-sorting steps in PResERV implicitly favors isolating just those antimutator alleles with no trade-off in terms of a reduced growth rate. However, there are potential risks and pitfalls in any directed evolution approach. Selection will yield a reduction in the dominant type of mutation for a particular reporter gene and plasmid in a particular environment and host cell, but these improvements may not translate to other DNA constructs, growth conditions, or genetic backgrounds. In this study, we showed that there are consistent antimutator effects between two distinct ColE1-type plasmids with different reporter genes in the evolved *E. coli* strains. How the antimutator alleles isolated here behave in other environments needs to be further tested to ensure that they do not degrade performance in specific applications. In general, this risk can be mitigated by matching PResERV conditions as closely as possible to those relevant for applications of a strain (e.g., in an

industrial bioreactor) or by exposing cells to a variety of different environments during PResERV. It also remains to be seen whether the PResERV mutations would maintain their antimutator effects if they were engineered into other *E. coli* strain backgrounds that are of interest in biotechnology (e.g., BL21 for protein expression).

One critical consideration for applying PResERV is knowing what types of mutations will inactivate the reporter gene used for cell sorting. Certain DNA sequences contain mutational hotspots such that a specific deletion or frameshift dominates among the mutations found to inactivate a reporter gene because it occurs at a rate that is many orders of magnitude higher than the point mutation rate<sup>49</sup>. Transposons are the most prominent source of mutations that disrupt other engineered DNA constructs<sup>22,23,107</sup>. The presence of any type of dominant mutation in the fluorescent reporter gene will concentrate PResERV on isolating mutants that ‘solve’ that particular mechanism of failure. In this study, we purposefully used a GFP reporter plasmid that had been edited to remove sequence-based mutational hotspots<sup>22</sup>, so that we could recover mutants that reduced point mutation rates. Because transposon mutations can be completely eliminated by using ‘clean-genome’ strains that have these and other selfish DNA elements deleted from their genomes<sup>23,48,50,51</sup>, it would probably not be a very useful application of PResERV to employ it to find mutations that suppress their activity, at least in bacteria. Rather, we anticipate that PResERV is most useful for neutralizing point mutations, for which it is less obvious how to modify either the sequence of the DNA construct or a cell’s genome to improve genetic stability.

One of the main challenges for implementing the PResERV approach in other contexts, as opposed to with a high-copy plasmid in a bacterial cell, is that expression of the reporter gene used to monitor for mutations must impose a large, dominant fitness burden on the host cell. This ensures, first, that cells with mutations in the reporter gene

will arise and reach a high frequency within a reasonable number of growth cycles so that one can complete multiple sorting steps to enrich for antimutator variants. Second, mutations that inactivate the reporter gene that is being monitored will be competing within these populations with other categories of beneficial mutations that improve growth for unrelated reasons (e.g., adaptation to the growth media). If the burden of the reporter gene is too small, then those other mutations will be favored over mutations that change GFP fluorescence, unfocusing evolution from the objective of PResERV. A related challenge, illustrated by the *polA* mutation in this study, is that it may be difficult to guard against a gradual and subtle loss of fluorescence over time during the sorting procedure, which can lead to the enrichment of mutations in the plasmid that modify the expression or burden of the reporter gene with undesirable side-effects.

In summary, we showed that the PResERV directed evolution approach can isolate antimutator *E. coli* variants that exhibit reduced mutagenesis of ColE1-type plasmids. Since these high-copy plasmids are widely used in *E. coli* for cloning and recombinant protein expression, these or similar antimutator alleles may be broadly useful in biotechnology applications. Future applications of PResERV with the burdensome reporter gene encoded in the chromosome or on a plasmid with a different origin of replication, might enrich for host cell variants that have a higher fidelity for replicating other components of a bacterial genome. The PResERV approach could also potentially be applied to other cell types used for industrial bioproduction, such as yeast or Chinese hamster ovary cells, if suitable reporter genes for monitoring genetic stability can be devised for these systems. Despite decades of studying the mechanisms of DNA repair and replication, we do not know the fundamental physiological constraints that determine a lower limit on the mutation rates that could potentially be achieved by tuning or augmenting these processes. Ultimately, this overall strategy of lowering mutation rates to

arrest evolution promises to improve the foundations of synthetic biology so that cells engineered for any purpose will function more predictably and reliably.

## **MATERIALS AND METHODS**

### **Culture conditions**

*E. coli* was grown as 10 mL cultures in 50 mL Erlenmeyer flasks with incubation at 37°C and 120 rpm orbital shaking over a diameter of 1 inch unless otherwise noted. The Miller formulation of Lysogeny Broth (LB) was used (10 g/L tryptone, 5 g/L yeast extract, and 10 g/L NaCl). Media were supplemented with 100 µg/mL carbenicillin (Crb), 20 µg/mL chloramphenicol (Cam), 50 µg/mL kanamycin (Kan), 100 µg/mL rifampicin (Rif), and 1 mM isopropyl β-d-1-thiogalactopyranoside (IPTG), as indicated. Bacterial cultures were frozen at –80°C after adding glycerol as a cryoprotectant to a final concentration of 13.3% (v/v).

### **Strains and plasmids**

The progenitor strain (BW25113) of the Keio knockout collection<sup>55</sup> was transformed with pSKO4. This plasmid contains the redesigned I7101 (R0010+E0240) circuit, which was edited to remove unstable repeat sequences in a prior study by Sleight *et al.*, on the BioBrick cloning vector pSB1A2 backbone<sup>22</sup>. It is a ColE1 group plasmid with a pBR322 origin of replication. Plasmid pTEM-1.D254tag encodes TEM-1 β-lactamase with the codon for an amino acid at a surface-exposed position in the enzyme's structure at which multiple amino acid substitutions are compatible with enzyme function replaced with a

TAG stop codon<sup>63</sup>. pTEM-1.D254tag has a pBR322 origin of replication and additionally encodes the *rop* protein.

### **UV mutagenesis**

BW25113 cells containing pSKO4 were cultured overnight to stationary phase in LB-Crb. Then, these cultures were pelleted by centrifugation and resuspended in an equal volume of sterile saline. Eleven 120  $\mu$ l droplets of these cell suspensions were spotted on petri dishes and subjected to 27,500  $\mu$ J/cm<sup>2</sup> of 254 nm UV radiation in a UVP CL-1000 crosslinker. After UV exposure, 100  $\mu$ l from each droplet was combined and pelleted by centrifugation to collect  $\sim 2.5 \times 10^6$  surviving cells. These cells were inoculated into 10 mL of LB-Crb and grown to a final density of  $\sim 2 \times 10^9$  cells/ml. This mutagenized library was archived as a frozen stock.

### **PResERV directed evolution procedure**

All growth steps were conducted in 10 mL of LB-IPTG-Crb. We used 0.1 mL of the mutagenized library to found the experimental population. After overnight growth to saturation, we propagated the population through daily 1:1000 dilutions of saturated cultures into fresh media followed by regrowth. GFP expression was monitored using a BD Fortessa flow cytometer. Periodically, overnight cultures were diluted to  $\sim 2.5 \times 10^6$  cells/mL in HPLC grade water and stained with 150 nM of the nucleic acid dye SYTO 17 (Life Technologies). The GFP+ portion of the population was calculated as the percentage of SYTO 17 positive cells with at least the ancestral level of GFP fluorescence by flow cytometry. When fewer than 25% of cells in the population were GFP+, instead of a normal

transfer, the population was diluted to  $\sim 2.5 \times 10^6$  cells/mL in HPLC grade water and between  $\sim 4 \times 10^4$  and  $\sim 5 \times 10^5$  GFP+ cells were sorted into 10 mL of fresh LB-Crb using a BD FACS Aria IIIu. The SYTO 17 dye was found to decrease cell viability, so we did not add this counterstain in the sorting steps, at the cost of less efficient enrichment of GFP+ over GFP- cells.

### **Isolation of evolved cells and plasmids**

The evolved population was plated on LB-IPTG-Crb and six visibly GFP+ colonies were selected at random for further study. Each of these clonal isolates was grown overnight in LB-IPTG-Crb before isolating its plasmid and creating a frozen stock. To select for plasmid loss, we also diluted these cultures 1:1000 into media lacking Crb (LB-IPTG). After overnight growth, dilutions of the resulting cultures were then plated on LB-IPTG agar. GFP- colonies were patched onto LB-Crb agar to ensure the lack of fluorescence was caused by loss of plasmid rather than a GFP mutation. One colony which had been cured of its plasmid was selected for each of the original evolved clones and re-transformed with the wild-type plasmid. We were unable to cure one evolved strain (AER7) of the plasmid in this way. The wild-type BW25113 strain was separately transformed with the plasmids isolated from each of the six evolved clones. GFP decay experiments were carried out on the resulting eleven strains to examine them for evidence of increased evolutionary stability.

### **GFP decay curves**

For each of the strains tested, individual colonies were used to inoculate nine replicate *E. coli* populations. In order to more accurately estimate the number of cell doublings elapsed since the single-cell bottleneck, care was taken to include all cells in each colony in the first liquid culture by excising and transferring the piece of agar underneath and around each colony. Each population was then subjected to daily transfers under the same conditions as the PResERV experiment, while monitoring GFP fluorescence using SYTO17 staining and a BD Fortessa flow cytometer as describe above. For creating graphs of the percentage of cells remaining GFP+ over time, flow cytometry data were analyzed in R using the flowCore Bioconductor package (v1.42.3). Among the events exhibiting a SYTO17 signal, cells were classified as GFP+ if they were above a signal intensity threshold that was set based on the distribution of fluorescence values observed for the wild-type strain-plasmid combination in that experiment. For graphing and comparing the initial GFP fluorescence in each strain, median intensity values for the GFP+ subpopulation were log2 transformed before performing statistical analyses.

### **Genome sequencing**

DNA was extracted from stationary phase *E. coli* cultures using the PureLink Genomic DNA Mini kit from Life Technologies. Purified DNA was fragmented using the Covaris AFA system, and samples were prepared using the NEBNext DNA Library Prep Reagent Set for Illumina kit from New England Biolabs. An Illumina HiSeq 2500 was used to generate  $2 \times 125$  paired-end reads from each sample at The University of Texas at Austin Genome Sequencing Analysis Facility (GSAF). FASTQ files have been deposited in the

NCBI sequence read archive (SRP090775). Mutations in each of the evolved strains were predicted by using *breseq* (version 0.32.0a)<sup>108</sup> to compare the Illumina reads to the *E. coli* BW25113 reference genome (GenBank:CP009273.1). Several genetic differences between this reference sequence and all four sequenced samples were assumed to have existed in the ancestral *E. coli* strain used to initiate the PResERV experiment and are not reported.

### **Strain reconstruction**

Donor strains from the i-Deconvoluter library<sup>109</sup> were used to revert the three candidate evolved alleles we tested in two evolved clones (AER8 and AER12) back to wild type sequences using P1 transduction as previously described<sup>110</sup>, except that only 2 µl of lysate was used in each transduction. Lysate from SMR20954, SMR20794, and SMR20838 was combined with the appropriate evolved strain with a mutation in *polA*, *polB*, or *rne*, respectively, and plated on LB agar plates containing Kan. Resultant colonies were screened for correct replacement via Sanger sequencing before FLP recombinase was used to remove the linked Kan<sup>R</sup> cassette used for selection of transductants as previously described<sup>111</sup>. The strain with *rne* reverted to the wild-type allele was subjected to a second round of P1 transduction to also revert the *polB* mutation, and the Kan<sup>R</sup> cassette was again removed via FLP recombination.

### **Mutation rate measurements**

Luria-Delbrück fluctuation tests were carried out to measure mutation rates<sup>112</sup>. For plasmid mutation rates, strains cured of plasmid pSKO4 were transformed with plasmid pTEM-1.D254tag after making any genetic modifications to revert evolved alleles. Cultures were



grown in LB-Cam to select for retention of this plasmid and plated on LB-Cam agar additionally supplemented with 500 µg/ml Crb to select for mutants. Thus, this assay measures the aggregate rate of all mutations that revert this stop codon to a permitted sense codon. Cells with the original, unmutated pTEM-1.D254tag plasmid were somewhat resistant to Crb, presumably due to some translational readthrough of the stop codon inserted into the β-lactamase gene on this high-copy plasmid. This background resistance is why an unusually high concentration of Crb (five times the level normally used to select for plasmid retention in this strain) was necessary to select for Crb<sup>R</sup> mutants. For chromosomal mutation rates, LB agar containing 100 µg/ml rifampicin (Rif) or 60 µg/ml d-cycloserine (DCS) was used for the selective conditions. Rif resistance requires specific point mutations in the *rpoB* gene<sup>113</sup>. In minimal media DCS resistance requires a loss-of function mutation specifically in the *cycA* gene<sup>114</sup>, but mutations in additional targets may also be possible in the rich LB media used here.

Mutation rates were determined by taking an overnight culture of a strain and transferring ~1,000 cells from a dilution in sterile saline to each separate fluctuation test culture. After growth of the replicate cultures to saturation, the entire volume of each one was either plated on a selective LB agar plate or a dilution was plated on a non-selective LB agar plate. For Crb<sup>R</sup> plasmid and Rif<sup>R</sup> chromosomal mutation rates comparing wild-type, evolved, and reconstructed strains, we used 0.2 mL cultures in 18×150 mm test tubes containing non-selective media and incubated these cultures with orbital shaking. A total of 6 non-selective and 48 or 12 selective plates were used for plasmid and chromosomal mutations, respectively. For DCS chromosomal mutation rates, 1.0 mL LB cultures and 12 selective media plates were used. Additionally, only a portion (25 µl) of these cultures were plated on the selection LB agar. For comparing Crb<sup>R</sup> plasmid mutation rates of wild-type and evolved clones (AER8 and AER12), we conducted three separate sets of fluctuation

tests using different growth formats. In the first, we grew 1 ml cultures in test tubes with orbital shaking and used 4 nonselective and 12 selective plates for each strain. In the second, we grew 200 µl cultures in test tubes with orbital shaking and used 12 nonselective and 48 selective plates for each strain. In the third, we grew 200 µl cultures in 96-well deep well microplates with no shaking and used 12 nonselective and 51 selective plates for each strain. In all cases, the liquid cultures in nonselective media were grown for 24 h, and mutant colonies on the selective plates were counted after 48 h of incubation. Colony counts on selective and non-selective plates were used to estimate mutation rates using the rSalvador R package (v1.7)<sup>115</sup>. We used its likelihood ratio methods for calculating confidence intervals and the statistical significance of differences between mutation rate estimates for two strains.

### **Plasmid copy number determination**

Cells containing the pTEM-1.D254tag plasmid were revived in 5 mL LB-Cam with overnight growth from frozen stocks. These cultures were diluted 1000-fold into 10 mL of fresh LB-Cam and 1 mL of each culture was harvested when its growth reached exponential phase ( $OD_{600} \sim 0.5$ ). Three cell pellets from different biological replicates were collected for each strain. Mixed plasmid and chromosomal DNA was isolated using the PureLink Genomic DNA Mini Kit (Invitrogen). Total DNA concentrations in these samples were determined using a Qubit Fluorometer (ThermoFisher).

Primer pairs for qPCR were designed to amplify products from either the antibiotic resistance gene (*cat*) in the pTEM-1.D254tag plasmid or the *ftsZ* gene in the *E. coli* chromosome. The *cat* primers were 5'-GTGAGCTGGTGATATGGGATAG and 5'-CCGGAAATCGTCGTGGTATT. The *ftsZ* primers were 5'-

GCAAGGTATCGCTGAACTGA and 5'-CGTAGCCCATCTCAGACATTAC. For each DNA sample, separate amplification reactions with each of the two primer pairs were conducted using Power SYBR Green PCR Master Mix reagents and a ViiA7 Real-Time PCR System (ThermoFisher). These reactions were performed in 96-well PCR plates with a final volume of 15  $\mu$ L per well and 500 nM of each primer.

Standard curves for plasmid and chromosomal DNA were used to calculate the absolute concentrations of each type of DNA from Ct values<sup>116</sup>. The plasmid standard curve was constructed by amplifying pTEM-1.D254tag plasmid DNA isolated using the PureLink Quick Plasmid Miniprep Kit (Invitrogen). The chromosomal DNA standard curve was made by amplifying *E. coli* DNA isolated using the PureLink Genomic DNA Mini Kit (Invitrogen) from cells without a plasmid. Each curve consisted of a series of 10-fold dilutions of template DNA. Plasmid copy number was calculated by averaging values from technical replicates for each biological replicate and then using the standard curves to estimate its plasmid:chromosome ratio. Graphing and statistical analyses were performed using log2 transformed values of copy number estimates for each biological replicate. The mean copy number estimated for the pTEM-1.D254tag plasmid in the wild-type BW25113 strain was 410.

### **Scaling of apparent mutation rates with plasmid copy number**

We used numerical simulations to examine how the apparent per-cell mutation rates estimated from our fluctuation tests with the pTEM-1.D254tag plasmid would be expected to scale if there were a change in the copy number of this plasmid in an evolved *E. coli* strain. Python scripts for performing the simulations are available online (<https://github.com/barricklab/plasmrs>). In these simulations, a population begins with a

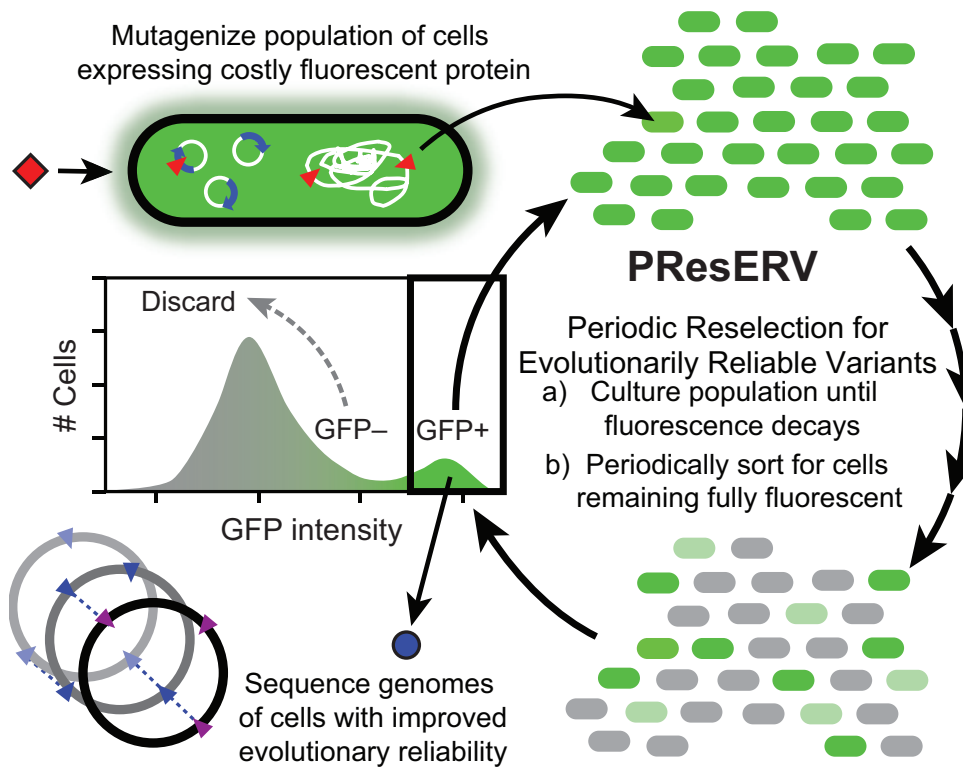
single wild-type cell that contains a set number of copies ( $N_p$ ) of the mutational reporter plasmid. The population growth process is modeled by iteratively picking a random cell from the population to divide and replacing this cell with its two daughter cells. When a cell divides, its complement of plasmids is replicated by iteratively picking a random plasmid from the current population of plasmids in the cell to copy until there are a total of  $2N_p$  plasmids. Each time a wild-type plasmid replicates there is a chance ( $\mu_p$ ) that the new copy is a mutant plasmid that has restored the  $\beta$ -lactamase reading frame. The resulting collection of  $2N_p$  plasmids, including any mutant plasmids that may have replicated or have newly arisen during division of this cell, are randomly allocated such that each daughter cell inherits exactly half of the plasmids. After cells divide enough times to reach a final population size ( $N$ ), the number of mutant cells that would yield colonies on Crb agar ( $N_m$ ) is counted as the number of cells that contain at least a minimum number of mutant plasmids needed to yield a resistant colony ( $N_r$ ). After one hundred replicate cultures were simulated for each condition, we estimated the apparent mutation rate per cell ( $\mu$ ) from the observed distribution of mutant colony counts per culture ( $N_m$ ) and the total number of cells per culture ( $N$ ) using the rSalvador R package, in the same way that we analyzed experimental data.

We specifically used these simulations to examine how the reduction in plasmid copy number observed in the AER12 PResERV strain with the *polA* mutation would be expected to impact its apparent mutation rate if the evolved strain maintained the same per-plasmid mutation rate as the wild-type BW25113 strain. Because it was not computationally feasible to simulate *E. coli* populations as large as those used in our actual fluctuation tests (with  $\sim 5.0 \times 10^7$  to  $\sim 2.5 \times 10^9$  cells) we performed simulations that scaled the apparent per-cell mutation rate ( $\mu$ ) upward such that equivalent values of the expected number of antibiotic resistant mutant cells per culture ( $m = N\mu$ ) were reached with smaller

values of  $N$ . We performed four different sets of simulations matching the  $\mu$  and  $m$  parameter combinations for the wild-type BW25113 strain that we observed in fluctuation tests in the four different experimental blocks shown in the results section that compared the plasmid mutation rates of the wild-type and AER12 strains. These combinations were:  $\mu = 4.69 \times 10^{-8}$  and  $m = 104$ ;  $\mu = 2.89 \times 10^{-7}$  and  $m = 27.7$ ;  $\mu = 2.36 \times 10^{-8}$  and  $m = 1.53$ ; and  $\mu = 2.14 \times 10^{-7}$  and  $m = 20.5$ .

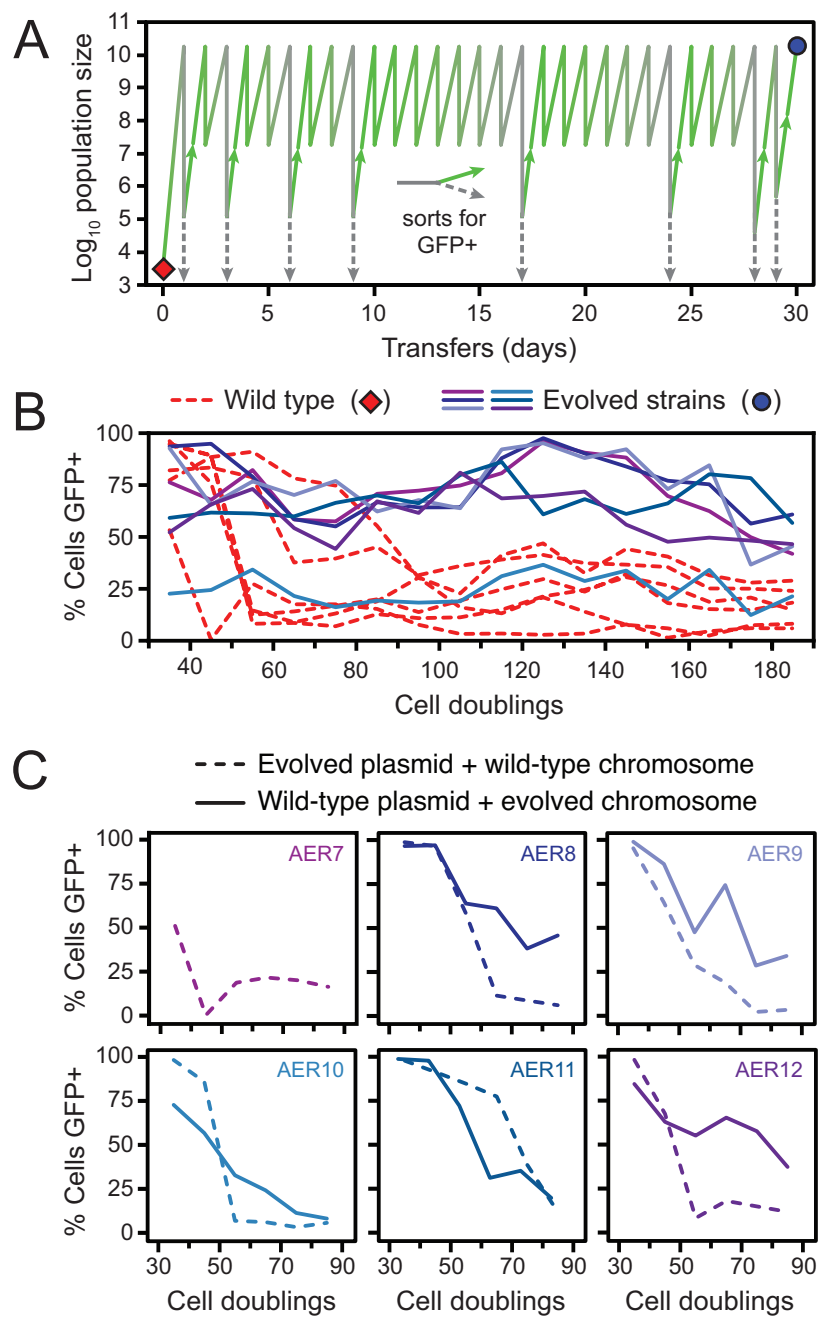
For each set of simulations, we first determined an underlying per-plasmid mutation rate ( $\mu_p$ ) that matched the experimentally measured apparent per-cell mutation rate ( $\mu$ ) to within 5% by performing a series of simulations using  $N = 3.2 \times 10^4$  cells and  $N_p = 410$  plasmids per cell. We repeated this procedure for each of three different values of  $N_r$  (1, 3, and 10). Then we performed five new simulations with each  $\mu_p$  value corresponding to a different  $N_r$  across five different values of  $N$  ( $10^4$ ,  $1.8 \times 10^4$ ,  $3.2 \times 10^4$ ,  $5.6 \times 10^4$ , and  $10^5$ ). Finally, we performed new simulations at these same fifteen combinations of  $\mu_p$ ,  $N$ , and  $N_r$  with  $N_p = 185$  and all other parameters left unchanged, in order to determine what the apparent mutation rate would have been in the fluctuation tests if plasmid copy number had decreased without any change in the underlying plasmid mutation rate. For each set of five pairs of simulations differing only in  $N_p$ , we calculated  $R$ , the ratio of the apparent mutation rate for  $N_p = 185$  to that for  $N_p = 410$ . We found no significant dependence of the logarithm of  $R$  on the logarithm of  $N$  across the range of values tested in any of these sets ( $p > 0.05$ , Bonferroni-corrected test for a nonzero slope in a linear regression model), justifying our inversely proportional rescaling of  $N$  and  $\mu$  for the purpose of making the simulations feasible. We further found that the values of the logarithm of  $R$  within each of the three sets of simulations at a fixed  $N_r$  that varied  $\mu$  and  $N$  were indistinguishable from one another ( $p > 0.05$ , Kruskal-Wallis tests). Therefore, we report only the overall mean of

the log-transformed  $R$  values for each set of 20 simulations at the same  $N_r$  and a bootstrap confidence interval on this statistic constructed from 100,000 resampled sets.



**Figure 2.1. Periodic Reselection for Evolutionarily Reliable Variants (PResERV) method**

PResERV begins with a population of cells expressing GFP to such a high level that it imposes a significant fitness burden. After mutagenesis, the population is cultured through enough cell doublings that mutants with reduced GFP expression arise and outcompete other cells. Periodically, the population is sorted to retain only those cells that remain as fluorescent as the original strain, enriching for mutant host cells with reduced mutation rates or a lower fitness cost for GFP expression. Once the evolutionary stability of GFP expression increases, fluorescent cells are isolated and their genomes are sequenced to identify and characterize the genetic changes that contribute to this improvement. Regrowth of cells during PResERV implicitly selects for only those mutants that achieve improved genetic stability without introducing any trade-offs that significantly reduce cellular growth rates.



**Figure 2.2. PResERV applied to an *E. coli* plasmid**

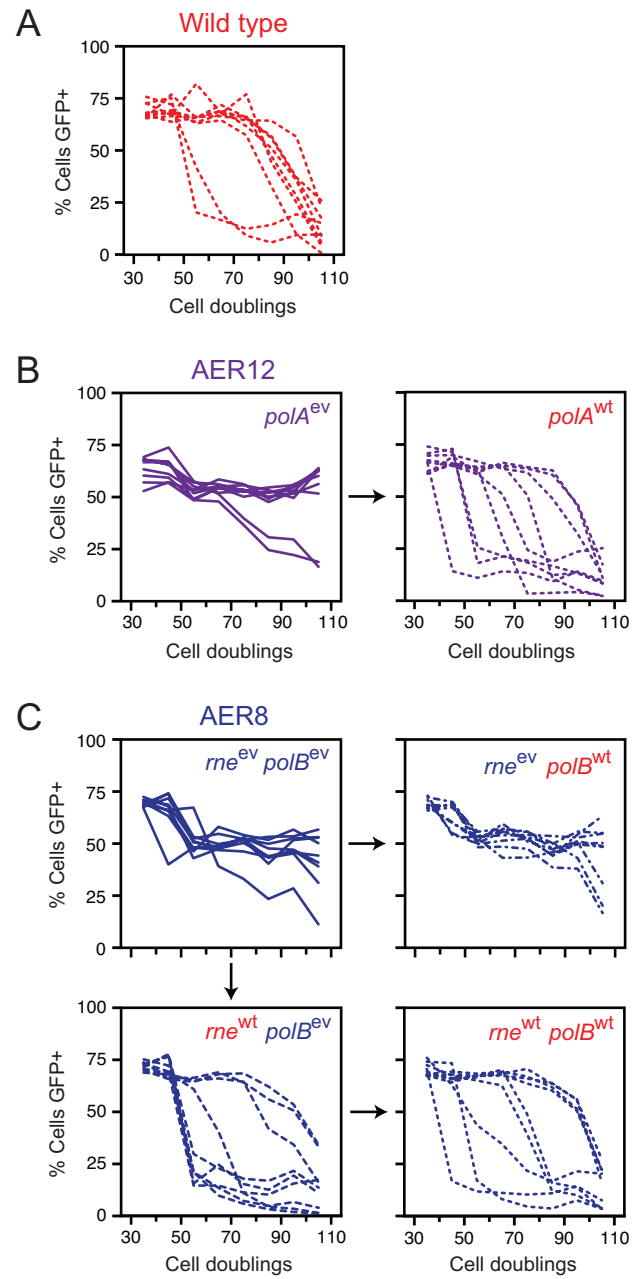


*Figure 2.2 continued* (A) Propagation and sorting regimen used to perform PResERV on an *E. coli* population in which GFP was expressed from plasmid pSKO4, a high-copy plasmid with a pBR322 origin of replication. The red diamond denotes the wild-type strain that was UV-mutagenized prior to beginning PResERV. Dashed grey and solid green bifurcating lines show when the population was sorted to retain fully fluorescent cells (GFP+). The blue circle indicates when fluorescent clones were isolated and sequenced. (B) Populations initiated from six different clones isolated at the end of the PResERV evolution experiment (blue and purple solid lines) were allowed to evolve alongside six replicates of the non-mutagenized, wild-type *E. coli* strain-plasmid combination (red dashed lines). Cells were considered GFP+ if they maintained a fluorescent intensity as measured by flow cytometry that was above a threshold level that was kept constant across all tested strains. (C) For each evolved PResERV strain, its plasmid was isolated and transformed into the wild-type strain containing no plasmid (dashed lines), and the evolved strain was cured of its plasmid and re-transformed with the wild-type pSKO4 plasmid (solid lines). Populations initiated from these strains were propagated and monitored as in B. The stability of AER7 re-transformed with the wild-type plasmid was not determined because of difficulty curing the evolved plasmid from this strain. In panels B and C, the same colors are used for each PResERV strain. In both experiments, the percentage of GFP+ cells was measured by flow cytometry after the first 35 cell doublings, corresponding to growth of the initial culture from a single cell on an agar plate, and then every 10 cell doublings afterward, corresponding to regrowth after daily subculturing steps that used a 1:1000 dilution into fresh media.

Position	Mutation	Evolved Strains				Annotation	Gene	Description
63,992	G→A	AER7	AER8	AER9		H597Y (CAT→IAT)	polB ←	DNA polymerase II
64,108	G→T	AER7	AER8	AER9		S558* (TCG→TAG)	polB ←	DNA polymerase II
516,812	A→G				AER12	E314G (GAA→GGA)	ybbP →	Putative ABC transporter permease
645,462	A→T	AER7	AER8	AER9		L162H (CTC→CAC)	citE ←	Citrate lyase (beta) subunit
1,139,159	A→G	AER7	AER8	AER9		L222S (TIA→TCA)	rne ←	Ribonuclease E
1,240,803	2 bp→AA	AER7	AER8	AER9		coding (188-189/441 nt)	ycgY →	Uncharacterized protein
1,444,811	G→A	AER7	AER8	AER9		L266L (CTG→ITG)	tynA ←	Tyramine oxidase, Cu-requiring
1,771,323	G→A	AER7	AER8	AER9		D494N (GAT→AAT)	ydiF →	Putative acetyl-CoA:acetoacetyl-CoA transferase
2,196,845	G→A				AER12	G1030R (GGG→AGG)	yehl →	DUF4132-containing protein
3,368,284	G→A	AER7	AER8	AER9		L19L (CTG→CTA)	dcuD →	Putative transporter
3,452,455	C→T				AER12	S253F (TCC→TTC)	gspE →	General secretory pathway component, cryptic
4,042,525	C→T				AER12	H734Y (CAC→IAC)	polA →	DNA Polymerase I
4,166,148	A→G			AER9		I93V (ATC→GTC)	tufB →	Translation elongation factor EF-Tu 2
4,327,401	IS1 (-) +8 bp			AER9		coding (339-346/762 nt)	adiY ←	Transcriptional activator, adi system

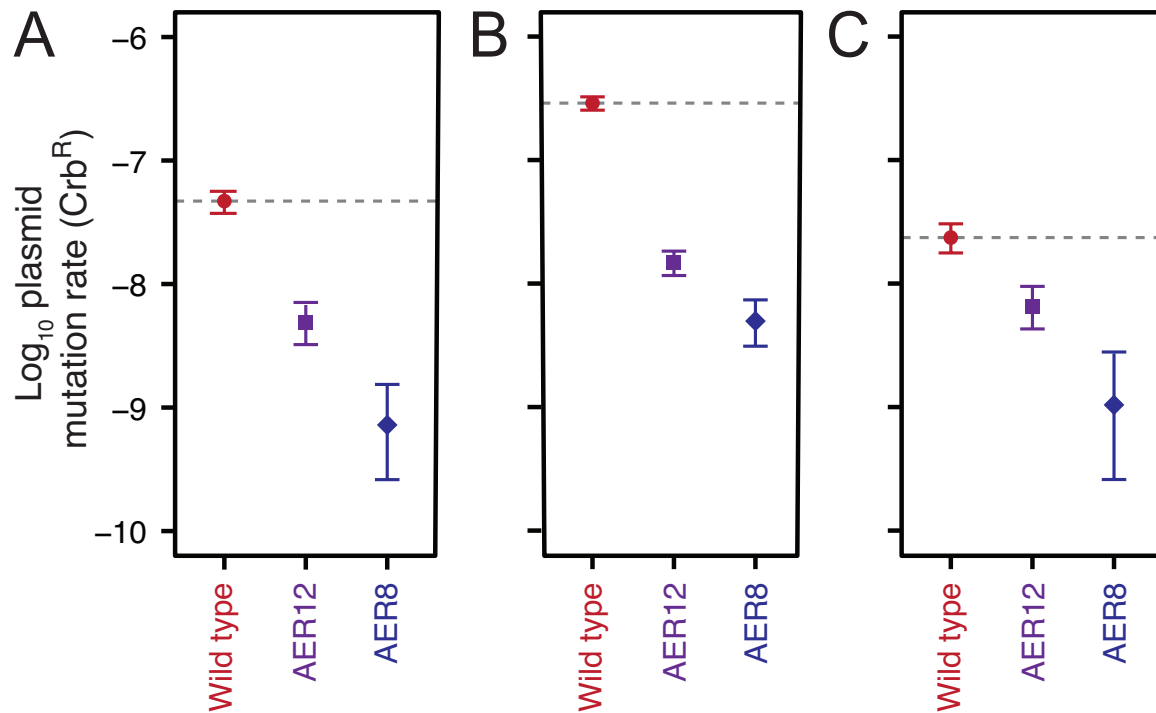
### Figure 2.3. Mutations in PResERV strains

The genomes of four evolved strains were re-sequenced to identify mutations that accumulated during the directed evolution procedure. The position column shows the coordinate of the first affected base pair defined relative to the *E. coli* K-12 BW25113 genome (GenBank:CP009273.1). The mutation column shows base changes on the top strand of the genome, except for the IS1 element in AER9 that inserted in the reverse direction and duplicated bases 4,327,401–4,327,408 at the target site on each side of the new IS copy. The annotation column shows the amino acid changes and codon changes caused by single-base substitutions or the locations of bases affected within a gene for other mutations. The gene column includes arrows showing the genomic strand on which each mutated gene is located.



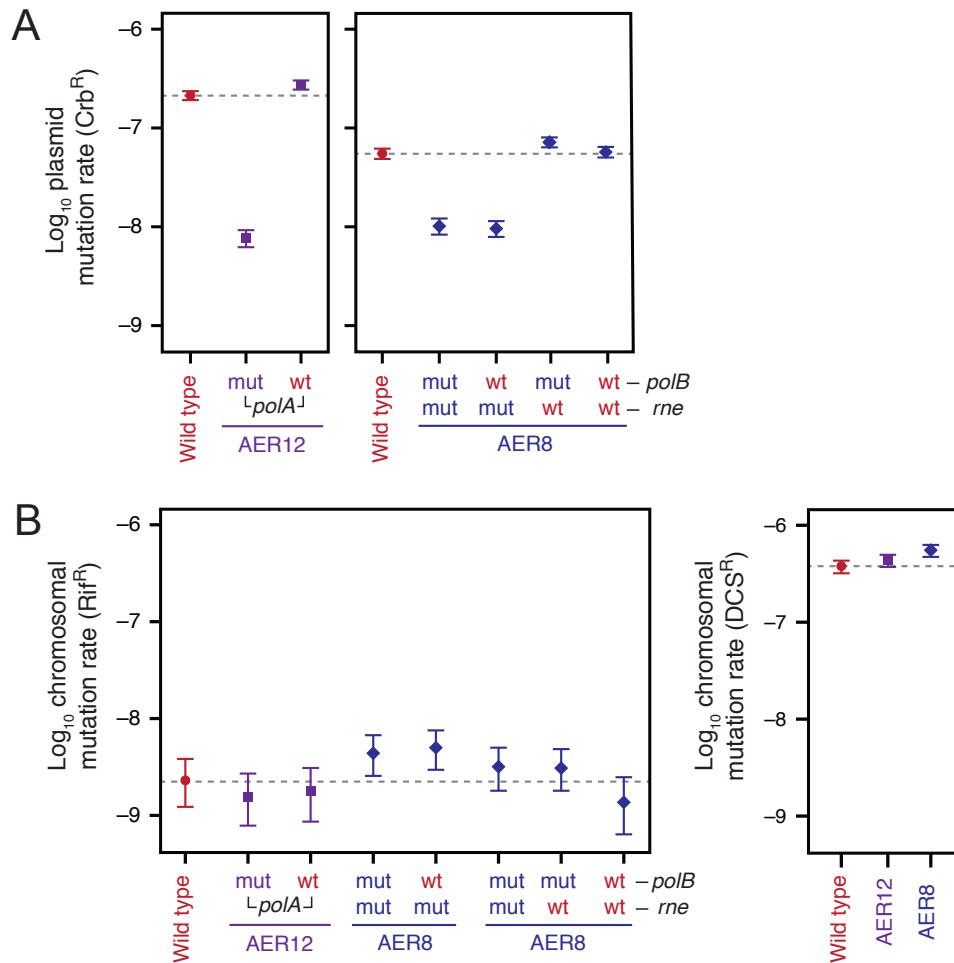
**Figure 2.4. Evolutionary stability and mutation rates in PResERV and reconstructed strains**

*Figure 2.4 continued* (A) Wild-type strain. (B) Evolved strain AER12 and a derived strain with its evolved *polA* allele reverted to the wild-type sequence. (C) Evolved strain AER8 and derived strains with its evolved *polB* and *rne* alleles reverted, singly and in combination. In each panel, the strains being tested were first transformed with plasmid pSKO4. Then, nine independent populations were initiated from single colonies of each strain. The prevalence of GFP-expressing cells within each population was monitored by flow cytometry over multiple daily serial transfers. Evolved strains are shown with solid lines. Dashed lines indicate that a strain contains one or more wild-type alleles, as indicated in red type. Measurements were made after the first 35 cell doublings and then every 10 cell doublings thereafter.



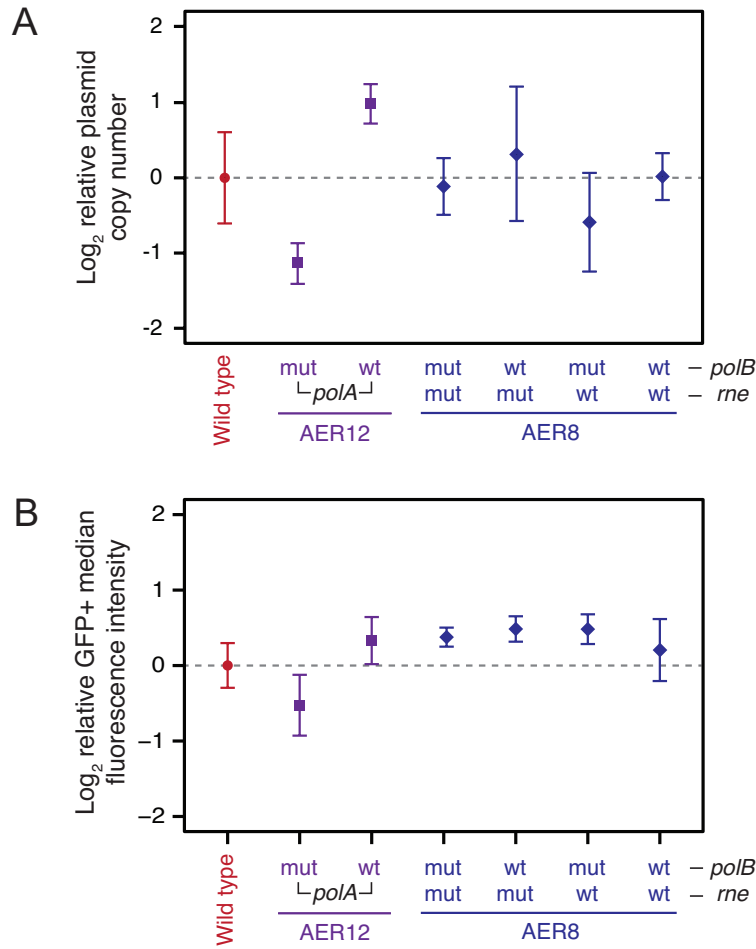
**Figure 2.5. Plasmid mutation rates in PResERV strains**

Mutation rates to carbenicillin resistance ( $\text{Crb}^R$ ) due to reversion of a stop codon in the TEM-1.D254tag plasmid were measured using Luria-Delbrück fluctuation tests. Wild type and the two focal evolved strains were compared in three experiments under different growth conditions: (A) in 1 ml cultures in test tubes incubated with orbital shaking (B) in 200  $\mu\text{l}$  test cultures in test tubes incubated with orbital shaking, and (C) in 200  $\mu\text{l}$  cultures incubated in 96-well microplates with no shaking. Each experiment included the wild-type *E. coli* strain for comparison (dashed horizontal lines). Error bars are 95% confidence intervals.



**Figure 2.6. Plasmid and chromosomal mutation rates in PResERV and reconstructed strains**

(A) Mutation rates to carbenicillin resistance (Crb<sup>R</sup>) due to reversion of a stop codon in the TEM-1.D254tag plasmid were measured using Luria-Delbrück fluctuation tests. Wild-type and evolved strains with mutant *polA*, *polB* and *rne* alleles (mut) reverted to wild-type sequences (wt), individually or in combination, were examined. Strains related to the evolved clone with a *polA* mutation (AER12) were tested in one experiment (left panel), and strains related to the evolved clone with *polB* and *rne* mutations (AER8) were tested in another experiment (right panel). (B) Mutation rates to rifampicin resistance (Rif<sup>R</sup>) (left panel) and d-cycloserine resistance (DCS<sup>R</sup>) (right panel) were measured using Luria-Delbrück fluctuation tests. Both of these resistance phenotypes require mutations in genes located on the *E. coli* chromosome. Each experiment included wild-type *E. coli* for comparison (dashed horizontal lines). Error bars are 95% confidence intervals.



**Figure 2.7. Plasmid copy number and GFP fluorescence in PResERV and reconstructed strains**

(A) Plasmid copy number for wild-type, evolved, and reconstructed strains determined by qPCR. Wild-type and evolved strains with mutant *polA*, *polB* and *rne* alleles (mut) reverted to wild-type alleles (wt), individually or in combination, were tested. The horizontal dashed line indicates the estimated copy number in the wild-type strain. Error bars show the standard error of the mean on log-transformed values from three biological replicates. (B) Initial GFP fluorescence of wild-type, evolved, and reconstructed strains as measured by flow cytometry. The median per-cell fluorescence intensity of the GFP+ subpopulation of cells was determined for each of nine replicate cultures immediately after outgrowth in liquid culture (after ~35 cell doublings). The graphed values are the log-averaged values of these medians. Error bars are 95% confidence intervals calculated assuming the logarithms of the medians are normally distributed. The horizontal dashed line shows the value for the wild-type strain.

### **Chapter 3: Reduced plasmid mutation rates in *Escherichia coli* evolve by mitigating the generation of reactive oxygen species through mutations that inactivate TCA cycle enzymes**

#### **ABSTRACT**

Biological systems are innately ephemeral, which makes them challenging to engineer. Living cells are able to rapidly evolve, an especially useful strategy when affronted by stressful conditions. Engineered cells contain synthetic devices, encoding a desired function, that are metabolically costly because their expression will sequester cellular resources and cause a burden on fitness. Therefore, there is strong selection for engineered cells that inactivate the synthetic device in order to alleviate the fitness cost. We have developed a directed evolution method, PResERV, in which antimutator strains are evolved and due to their reduced mutation rates, have a lower probability of inactivating costly synthetic devices. Here, we evolved and characterized a series of *E. coli* antimutator strains. We determined the genetic basis of the antimutator phenotype to be largely caused by mutations in two genes related to central carbon metabolism, *sucD* and *sdhA*. Furthermore, these antimutator alleles function to mitigate intracellular oxidative stress, leading to less DNA damage, and the observed antimutator phenotype. The *sucD* and *sdhA* antimutator alleles are highly conserved, and were therefore found to reduce mutation rates in another *E. coli* strain, BL21. Notably, the *sucD* and *sdhA* alleles can be combined with a previously identified antimutator allele in *rne*, Rnase E, to reduce mutation rates even further. Our work provides genetically stable antimutator strains for scientists to tackle complex engineering problems and, describes a mechanism involved in the antimutator phenotype that will guide future construction of antimutator strains.



## INTRODUCTION

Microbes are tiny factories. Their abilities have been successfully and unsuccessfully co-opted by scientists for decades. The most quotidian examples of successfully engineered biologics are insulin, recombinant antibodies, the anti-malarial drug artemisinin, and more recently, the HPV vaccine. Biology can be engineered to address a wide variety of societal issues and the limit, should lie within the imagination of the scientist. Currently, there exist efforts to engineer biology for diverse applications in areas such as space exploration, biomaterials, plant and human probiotics, and diagnostic platforms. So, why haven't more of these technologies transitioned as commonplace in our daily lives? Biological systems are inherently unpredictable, making them challenging to engineer.

The difficulty exists in an idiosyncrasy common in all living cells – evolution. In order to engineer cells for a given purpose, synthetic devices encoding a desired function need to be inserted into a host cell and their expression is dependent on the host machinery. The host cell is not adapted to manage burdensome expression of this additional, foreign DNA and its intracellular stoichiometry becomes unbalanced. This stress favors mutants that can alleviate the burden by evolving, and subsequently breaking expression of the synthetic device. Unpredictability stems from these inactivating mutations which in turn, hamper engineering efforts. To address this problem, we need to properly redesign synthetic DNA constructs and host cells to resist evolution<sup>11,45,47</sup>.

Evolutionary failure modes consist of several common themes. Some, such as recombination between homologous sequences and slippage errors due to simple sequence repeats, are easy fixes and can simply be avoided by omitting these sequences from the design of the synthetic device<sup>22,47,49</sup>. Other failure modes require extensive engineering and are independent of diligently crafted constructs. Host organisms contain active mobile insertion sequences (IS) that will readily eliminate engineered functions in a cell<sup>23</sup>. A

method to attenuate this problem is to build IS-free versions of host strains. IS-free hosts are better able to maintain expression of synthetic constructs and have reduced mutation rates<sup>39,48,50,107,117</sup>. However, these hosts do not help mitigate evolutionary failure by single base-pair changes, an even more challenging problem. To address this issue, we have developed a directed evolution method (Periodic Reselection for Evolutionarily Reliable Variants, PResERV) to evolve genetically stable, antimutator strains. These strains are resistant to evolutionary failure due to their reduced mutation rates and resulting lower probability of mutations that will inactivate an engineered function.

We previously performed a pilot PResERV experiment in *E. coli*, in which we identified antimutator alleles in *polA*, *polB*, and *rne*<sup>118</sup>. Here, we applied PResERV on a larger scale, using a similar experimental setup to identify more mechanisms linked to the antimutator phenotype. We evolved a series of antimutators via PResERV and determined that the genetic basis of their reduced mutation rates was mainly due to mutations in genes involved in central carbon metabolism, *sucD* and *sdhA*. Further characterization of these antimutator alleles demonstrated that they are complete inactivations of their respective protein complexes and ultimately serve to alleviate intracellular oxidative stress. Additionally, we found that the *sucD* and *sdhA* alleles reduce mutation rates in another *E. coli* strain, BL21, and function in combination with the *rne* allele, from the initial PResERV experiment, to lower mutation rates even further.

## RESULTS

### Directed evolution via PResERV

To isolate new antimutator variants of *E. coli*, we employed an improved version of the Periodic Reselection for Evolutionarily Reliable Variants (PResERV) method

previously developed by our lab<sup>118</sup>. PResERV is a directed evolution approach that uses cell sorting to enrich for mutant cells that have a greater chance of continuing to express a burdensome fluorescent reporter gene over many generations of growth. Mutants that maintain long-term expression of the synthetic construct may have increased genetic stability due to evolving reduced mutation rates. In this study, we implemented sorting conditions that better enrich for fluorescent versus nonfluorescent cells by using a nonlethal lipophilic membrane dye and reducing the flow rate. We performed PResERV on five independently mutagenized and propagated populations of *E. coli* BW25113 harboring the plasmid pSKO4, a high-copy pBR322-derived plasmid expressing GFP. We performed an average of 10 sorts spread over ~350 generations (~30 days) of evolution. After evolution, we isolated 70 individual clones from each population to examine the time-course of GFP inactivation in the context of a single strain background. For follow-up studies and whole-genome sequencing, we selected three clones per population (designated A, B, and C) that demonstrated an increased ability to maintain GFP production compared to wild type.

### **Evolved strains have reduced mutation rates**

To determine if the improvement in the evolutionary stability of GFP expression was due to reduced mutation rates in the PResERV strains, we measured plasmid and chromosomal mutation rates in the 15 selected clones using Luria-Delbrück fluctuation tests (**Figure 3.1**). The assay for measuring plasmid mutation rates used the pBR322-derived plasmid pTEM-1.D254tag<sup>63</sup>. It encodes a non-functional TEM-1  $\beta$ -lactamase reporter gene that, when mutated, renders cells resistant to carbenicillin resistance. The  $\beta$ -lactamase mutational reporter gene is inactivated by a nonsense mutation in the coding region. Therefore, this assay specifically captures the rate of point mutations on the plasmid

that convert this codon back to any one of multiple sense codons compatible with protein function<sup>118</sup>. We found that 12 of the 13 tested PResERV-evolved strains exhibited plasmid mutation rates that were significantly lower than the wild-type strain ( $p\text{-adj} > 0.05$ , Bonferroni-adjusted p-value) (**Figure 3.1A**). The reduction in mutation rates was as large as 10-fold in some strains.

To determine whether there was a change in chromosomal mutation rates in the evolved strains, we performed fluctuation tests that selected for rifampicin resistant mutants (**Figure 3.1B**). Resistance to this antibiotic arises due to point mutations in the *rpoB*, in which ~100 single base-pair substitutions will result in rifampicin resistance<sup>113</sup>. Compared to the plasmid rates, there appeared to be a modest reduction in chromosomal mutation rates, in most evolved clones, though none of the clones were statistically significant ( $p\text{-adj} > 0.05$ , Bonferroni-adjusted p-value) (**Figure 3.1B**). This result is not surprising given that the GFP gene that was monitored for stability during PResERV was on a high-copy plasmid. Therefore, selection could only directly favor mutations that improved maintenance of this costly gene, which can be accomplished by reducing plasmid mutation rates.

### **Most PResERV strains maintain wild-type plasmid copy number and growth rates**

Plasmid copy number can potentially affect the apparent mutation rate measured in a reporter gene on a plasmid. For example, if there are half as many plasmids in a cell, then its apparent mutation rate will be reduced by approximately two-fold<sup>118</sup>. Therefore, we measured the copy number of our *ampR* mutation rate reporter plasmid in the PResERV strains to determine whether any of the evolved clones adapted a lower plasmid copy number during PResERV evolution (**Figure 3.2A**). We found only two strains to be

affected, 3C and 5C, and these were discarded from further study. These had ~60 and 7-fold reduced copy number, respectively.

One advantage of selection during PResERV is that there is enrichment for antimutator alleles that lack any deleterious effects on growth. Cells must compete for resources within a population and iterative cell sorts will favor mutants that are more abundant in the population. Thus, PResERV is designed to discard slow growing mutants. We found that the PResERV antimutators have similar growth doubling times as wild-type in rich and minimal media, confirming that their decreased mutation rates do not come at the expense of a reduced growth rate, at least in the environment used for the PResERV experiment (**Figure 3.2B**).

### **Mutations in the genomes of PResERV strains**

Given that PResERV involves directed evolution, our method allows us to explore new antimutator pathways that are non-obvious targets for reducing mutation rates. We found an average of six genomic mutations per clone and a total of 12 synonymous mutations altogether. There were no observed mutations in the pSKO4 plasmid. A subset of the whole-genome sequencing data showing genes hit by mutations in >2 strains and other mutations in genes with functions potentially related to mutation rates is summarized in **Table 3.1**. The most commonly mutated genes are *sucD*, *sdhA*, *aspA*, *hofC*, *ulaR*, and *frlA*. Many of these genes are major enzymes in central carbon and amino acid metabolism. In addition, there were mutations in several genes (*btuE*, *dinF*, and *yjjX*) involved in oxidative stress, which may be evidence of a causative mechanism for reducing mutation rates.

### Mapping the antimutator phenotype

To identify the alleles contributing to reduced mutation rates in the evolved strains we individually reverted 27 of the mutations in **Table 3.1** in at least one PResERV evolved strain that had the mutation. We used fluctuation assays to quantify the effect of the reverted allele on plasmid mutation rates. Those alleles that partially or fully restored the wild-type mutation rate were classified as causative. For most of the PResERV strains we determined that the causative alleles were in these key genes – *sucD*, *sdhA*, *adhE*, *dinF*, and *ygcL* – and two of these genes, *sucD* and *sdhA*, were mutated in multiple PResERV strains. (**Figure 3.3A, B, and D**). These alleles were consistently causative, independent of the evolved strain background. The *sucD* allele (E98\*) was identified in all three strains from one population (1A, 1B, and 1C), indicating that this mutation was likely present in the starting library of mutants and was enriched during PResERV evolution. Conversely, there were four *sdhA* alleles (Y78L, N156Kfs, Q173\*, and Q509\*) identified in five evolved strains from different populations (**Table 3.1**). This suggests a high level of parallel evolution occurring during PResERV.

In some instances, we were unable to determine the causative mutations (**Figure 3.3C and E**). For example, we reverted two alleles in *aspA*, another frequently mutated gene, and one allele resulted to be causative in only a single clone (**Figure 3.3C**, clone 3A). In the other clones, either another mutation was found to be causative (**Figure 3.3B**, clone 2B) or the original evolved clone did not have a reproducibly low mutation rate across multiple fluctuation assays performed on different days (**Figure 3.3C**, clone 3B). Therefore, the evolved alleles in *aspA* did not appear to be consistently causative. Clones 3A and 3B did not contain any other candidate mutations and possibly, the causative mechanism lies in a cellular regulatory change not described by mutations in the genome. Similarly, clones from population five proved difficult to diagnose. Clone 5A also

contained only one mutation, in *sucC*, which when reverted, did not appear to be causative and clone 5B did not have a reproducibly low mutation rate (**Figure 3.3E**, clones 5A and 5B). Additionally, these two populations contained clones, 3C and 5C, that had problems with plasmid copy number (**Figure 3.2A**). Given these difficulties, our follow-up studies focused on those mutations that were bona fide causative.

We independently reconstructed the causative alleles in a wild-type background to discern whether the antimutator phenotype was contingent solely on the single allele and not any other mutations present in the background of the evolved strain. We found that the *sucD* and *sdhA* mutations alone, in a wild-type background, resulted in an antimutator phenotype (**Figure 3.4B**). The other alleles (*adhE*, *dinF*, and *ygcL*), which were previously identified as causative failed to reproduce the antimutator phenotype in a wild-type background and were not included in further study.

### **Loss of function mutations in *sucD* and *sdhA* reduce the *E. coli* basal mutation rate**

We focused further on mutations in two genes, *sucD* and *sdhA*, that encode subunits of key enzymes in the tricarboxylic acid (TCA) cycle and the electron transport chain (ETC). These genes are genomically located within a cluster, along with other TCA cycle genes, and the operon structure is illustrated in **Figure 3.4A**. Succinyl Coenzyme A synthetase (SucCD) catalyzes the conversion of succinyl-CoA to succinate and is the only TCA cycle step in which substrate level phosphorylation of ATP occurs<sup>119</sup>. It is a heterotetramer which consists of two  $\beta$  (SucC) and  $\alpha$  (SucD) subunits, and the  $\alpha$  subunit hosts the catalytic phosphorylation site<sup>119,120</sup>. The causative PResERV evolved allele, E98\*, is a nonsense mutation that likely renders the catalytic  $\alpha$  subunit and the whole protein complex inactive. The subsequent metabolic reaction converting succinate to fumarate is catalyzed by

succinate dehydrogenase (SdhCDAB). Succinate dehydrogenase is the only TCA cycle enzyme that also functions in the ETC and is therefore embedded in the membrane<sup>121</sup>. The SdhCDAB protein complex is a trimer composed of four subunits in which two are structural, containing transmembrane domains (SdhCD), and the other two harbor electron-shuttling cofactors, flavin and iron-sulfur clusters, that perform catalysis (SdhAB)<sup>122</sup>. The causative *sdhA* PResERV alleles consist of nonsense and frameshift mutations that expectedly result in a nonfunctioning *sdhA* subunit, which could lead to either a reduced or completely inactive enzyme complex.

To further study the genetic basis of the PResERV *sucD* and *sdhA* alleles, we asked whether the *sucD* and *sdhA* evolved alleles were loss-of-function mutations. To do this, we assayed the mutation rate of *sucD* and *sdhA* KEIO deletion mutants (**Figure 3.4B**). In the case of *sucD*, the evolved allele and the KEIO deletion mutant had similar mutation rates, indicating that the PResERV *sucD* allele is indeed a loss-of-function mutation. A similar situation is observed with *sdhA*. There are two classes of *sdhA* alleles based on the fold reduction in mutation rate. The first class, Y78L, results in a modestly reduced mutation rate whereas the second class, that are expectedly loss-of-function (N156Kfs, Q173\*, and Q509\*), have a marked antimutator effect. The *sdhA* KEIO deletion mutant has a similarly reduced mutation rate as this second class of *sdhA* alleles, which demonstrates that these alleles are deletions of *sdhA*. The first class allele, Y78L, likely renders an intermediate antimutator phenotype because the mutation does not result in a complete knockout of *sdhA*.



### **Further dissection of the antimutator roles of succinyl-CoA synthetase and succinate dehydrogenase**

As described in the previous section, SucD and SdhA are components of a larger protein complex. Since deletion of these subunits results in an antimutator phenotype, what effect will deletion of the other protein subunits have on the mutation rate? To address this, we measured the mutation rates of *sucC*, *sdhB*, *sdhC*, and *sdhD* KEIO deletion mutants (**Figure 3.4B**). Deletion of the other succinyl CoA synthetase subunit, SucC, did not reduce the mutation rate. The catalytic residue of succinyl CoA synthetase resides in SucD and therefore, deletion of *sucC* may result in an enzyme with reduced activity instead of a completely non-functional variant. The antimutator phenotype possibly relies on the latter. Similarly, deletions of *sdhB* and *sdhC*, but not *sdhD*, resulted in lower mutation rates. The *sdhB* and *sdhC* deletions may lead to a complete inactivation of the succinate dehydrogenase enzyme complex, whereas deletion of *sdhD* could have little effect on the overall enzyme activity. In sum, our data support that the PResERV *sucD* and *sdhA* alleles are knockouts of their respective protein subunits, rendering a complete inactivation of the entire enzyme complex.

### **Antimutator alleles function to alleviate DNA damage induced by ROS**

Reactive oxygen species (ROS) are toxic byproducts of aerobic cellular metabolism. The formation of these species, hydrogen peroxide (H<sub>2</sub>O<sub>2</sub>), superoxide (O<sub>2</sub><sup>-</sup>), and hydroxyl radical (OH<sup>-</sup>), engenders damage to a cell's basic building blocks – nucleic acids, proteins, and lipids<sup>123</sup>. Intracellular ROS is generated predominantly by enzymes in the TCA cycle and ETC<sup>124</sup>. Given that SucD and SdhA are key components in these pathways, we examined whether the antimutator *sucD* and *sdhA* alleles function to reduce intracellular ROS stress.

One method to study ROS stress is to induce its formation by using redox-cycling compounds such as paraquat. Redox-cycling compounds will diffuse into the cell and catalyze electron transfer from a redox enzyme to molecular oxygen. This reaction will generate hydrogen peroxide and superoxide, which can subsequently be converted into hydroxyl radicals<sup>125</sup>. Therefore, all three species are artificially induced within a cell based on the availability of electron-donating enzymes. Cells containing more of these enzymes will experience higher rates of lethality in the presence of paraquat. We assayed survival of the antimutator *sucD* and *sdhA* alleles after paraquat exposure (**Figure 3.5A**). The antimutator strains exhibited higher tolerance levels than wild-type and a control strain known to be more susceptible to redox-cycling drugs ( $\Delta zwf$ )<sup>126</sup>. These results suggest that the *sucD* and *sdhA* antimutator cells have lower levels of intracellular ROS.

To directly capture intracellular ROS concentrations, we exposed the antimutator strains to an oxidation-sensitive fluorescent probe (2',7'-dichlorofluorescein diacetate). This probe is activated intracellularly via oxidation by ROS and therefore, elevated concentrations of ROS will result in higher fluorescence levels. Our data demonstrate a significant decrease in median fluorescence, and therefore ROS levels, in the *sucD* and *sdhA* antimutator strains (Q173\* and Q509\*) as compared to wild-type and the  $\Delta zwf$  positive control ( $p < 0.05$ , Mann-Whitney *U* test) (**Figure 3.5B**).

Based on our data, we propose that by reducing ROS stress, a *sucD* or *sdhA* antimutator cell experiences less DNA damage and therefore low mutation rates. This mechanism ultimately leads to the increased stability of synthetic constructs.

### **PResERV alleles reduce mutation rates in other strains**

Since these enzymes are widely conserved in other organisms, we hypothesized that the PResERV alleles could be ported into the genome of another *E. coli* strain, primarily used for plasmid-based recombinant protein production, BL21<sup>127,128</sup>. The wild-type strain used in this PResERV study (BW25113) is a K-12-derived strain whereas BL21 originated from *E. coli* B<sup>129,130</sup>. *E. coli* K-12 and B isolates have highly similar genomes and differ in that BL21 has a rapid growth rate, low accumulation of acetate, and improved protein production due to the absence of proteases<sup>130,131</sup>. Despite these physiological differences, our K-12-isolated PResERV antimutator alleles successfully reduced the BL21 mutation rate (**Figure 3.6A**). This reduction in mutation rate will allow for increased stability of protein expression plasmids in BL21.

### **Total Antimutator Combination Organism (TACO)**

Our work characterizes mechanisms involved in genetic stability of synthetic constructs in order to engineer reliable, workhorse strains for biology. As a step towards this effort, we created Total Antimutator Combination Organisms (TACO) containing two antimutator alleles and assayed whether these alleles would function to reduce the mutation rates even further. We combined each *sucD* and *sdhA* allele with another antimutator mutation, *rne* (L222S), from a previous study<sup>118</sup>. Indeed, the double mutant TACO strains exhibited a further reduction in mutation rates when compared to the single mutants (**Figure 3.6B**). These engineering experiments are the nascent stages of future work. In principle, other versions of TACO can be created by combining different antimutator alleles in *E. coli* or other organisms of interest. These TACO strains can be utilized as new,

stable host organisms for biology and as examples of where the limits exist for reducing mutation rates.

## DISCUSSION

Our study describes a series of antimutator strains that were evolved to maintain expression of a costly, high-copy (pSKO4) via PResERV. These strains have varied plasmid mutation rates across strains/populations, and many of these are marked reductions (1.5 to 15-fold). In contrast, chromosomal rates tended to be uniform and only modest reductions were observed (1.1 to 2-fold). This antimutator phenotype is caused by mutations in two genes, *sucD* and *sdhA*, which are common targets in this PResERV experiment. More than half of the 15 antimutator strains contained mutations in these genes. In further dissecting this phenotype, we determined that *sucD* and *sdhA* are loss-of-function mutations which ultimately lead to inactivation of their associated protein complexes, succinyl-CoA synthetase and succinate dehydrogenase, respectively. Moreover, inactivation of these protein complexes results in reduced intracellular ROS levels which likely accounts for the observed reduction in mutation rates. Given ubiquity of this mechanism across organisms, we ported the *sucD* and *sdhA* antimutator alleles into another *E. coli* strain, BL21, and found that they again function to reduce mutation rates. Finally, we were able to reduce mutation rates further and build even more stable strains by combining antimutator alleles to engineer TACO. In generating a series of stable strains and delineating their causative mechanism, our work provides resources for scientists to better understand and engineer biology.

## Different mutational solutions to the same problem

We previously performed the first PResERV pilot experiment to evolve antimutators with the same costly plasmid, pSKO4, that was used in this work<sup>118</sup>. Notably, the causative mutations and proposed mechanisms found in this initial study did not overlap with those described here, despite the same costly plasmid used for evolution. The initial PResERV pilot found that mutations in *rne*, *polB*, and *polA* lead to an antimutator phenotype. These mutational targets suggest that pathways such as DNA/RNA metabolism and processing are mechanistically involved, which is unlike the ROS mechanism that is extensively discussed in this work. Two main possibilities underlie this difference. Firstly, there were several changes to the PResERV protocol that primarily involved using a different fluorescent dye to visualize the population, a slower flow rate to better separate the fluorescent cells, and the PResERV pilot experiment consisted of a single population whereas this work evolved five populations in parallel.

Secondly, the GFP decay genetic screen to identify putative antimutators from a whole population was unlikely saturated in both experiments. In the pilot experiment we selected six clones for characterization and in this work, we picked three clones per population. Since only a few clones were characterized in each study, we were unable to obtain a comprehensive idea of the genotypes present in the whole population and whether some of the mutations were present in both PResERV experiments. Despite this apparent dearth in characterizing antimutators, we were able to identify antimutator alleles that are non-obvious targets for genetic stability.

### **How do loss-of-function antimutator alleles in *sucD* and *sdhA* result in decreased ROS?**

Cells contain many intricate mechanisms for neutralizing oxidative stress. Namely, the most recognized systems are of transcriptional regulators SoxRS and OxyR, which respond to  $O_2^-$  and  $H_2O_2$ , respectively<sup>123,124</sup>. There also exist a myriad of strategies involving redirection of central carbon metabolism and these consist of two principal themes:

1. Detoxification of ROS by  $\alpha$ -keto acid TCA cycle intermediates
2. Modification of the intracellular NADPH:NADH ratio because NADH is a known pro-oxidant that magnifies cellular oxidative tension

These strategies are discussed below as they relate to either the *sucD* or *sdhA* antimutator alleles.

#### ***sucD***

As described in **Figure 3.4A**, SucD is the  $\alpha$  subunit of succinyl-CoA synthetase, and it converts succinyl-CoA to succinate. It contains the catalytic residue, His247, which is the site for autophosphorylation of the  $\alpha$  subunit and the very first step in the reaction mechanism<sup>132</sup>. Enzyme variants of SucD containing mutations that disrupt this site are non-functional<sup>133</sup>. The  $\beta$  subunit, SucC, is the nucleotide-binding domain responsible for the final part of the reaction mechanism that involves of substrate level phosphorylation to generate ATP<sup>119,134</sup>. Therefore, deletion of *sucC* results in an enzyme complex that retains the ability to convert to succinate, but is impaired in synthesis of ATP<sup>135</sup>. Given these phenotypes and our data showing that the KEIO *sucD* deletion has a reduced mutation rate (**Figure 3.4B**), we conclude that the *sucD* antimutator allele is completely inactivating succinyl-CoA synthetase and therefore, rerouting flux through the TCA cycle. A non-functional succinyl-CoA synthetase can result in decreased intracellular ROS due to two

metabolic changes within the TCA cycle – accumulation of  $\alpha$ -ketoglutarate ( $\alpha$ -KG) and increased flux through the glyoxylate shunt.

During oxidative stress,  $\alpha$ -ketoglutarate dehydrogenase becomes inhibited by ROS and its substrate,  $\alpha$ -KG will accumulate<sup>136</sup>.  $\alpha$ -KG is a powerful antioxidant, scavenging ROS and concomitantly producing a readily usable metabolite, succinate. This strategy is advantageous because it does not rely on the availability of NADPH and in fact, more intracellular NADPH is generated via upregulation of the NADPH-producing enzyme, isocitrate dehydrogenase<sup>137</sup>. In addition, NADH synthesis is limited by increased flux through the glyoxylate shunt, which is a known adaptation during oxidative stress<sup>138,139</sup>. These changes work to mitigate intracellular ROS via bypassing the main TCA cycle, generating fewer NADH molecules, and reducing flux through the ETC<sup>139</sup>. Furthermore, glyoxylate is another  $\alpha$ -keto acid, analogous to  $\alpha$ -KG, and will therefore serve to quench ROS resulting in production of formate<sup>138,140</sup>. Glyoxylate can also be enzymatically converted into glycine, a component of the major oxidative stress defense molecule, glutathione<sup>138</sup>. We hypothesize that these metabolic reconfigurations occur in a cell lacking succinyl-CoA synthase and therefore, mechanistically support the *sucD* antimutator phenotype.

### ***sdhA***

The enzyme complex of succinate dehydrogenase consists of two electron-shuttling subunits, SdhAB, and two membrane anchor proteins, SdhCD (**Figure 3.4A**). Succinate dehydrogenase performs two key functions. The initial part of the reaction mechanism, oxidation of succinate to fumarate, is performed by the cytoplasmic subunits, SdhA and SdhB. Following this conversion, the resulting electrons are transferred to the

transmembrane anchors, SdhC and SdhD, and are further transported through quinones to the rest of the ETC. In order for this cascade to occur, succinate dehydrogenase contains a series of redox cofactors, flavin adenine dinucleotide (FAD) and iron-sulfur clusters, to move electrons throughout the protein complex. Without this intact cascade, the redox potential of the enzyme is lowered and the initial oxidation of succinate to fumarate cannot occur. This precludes the movement of electrons from this reaction into the ETC<sup>121</sup>. Moreover, redox enzymes are known to produce ROS during normal catalysis by spontaneous enzyme autoxidation<sup>124</sup>. In succinate dehydrogenase, the FAD cofactor in SdhA is the primary site of autoxidation since it is predominantly solvent exposed<sup>141</sup>. Deletion of *sdhA* will prevent this autoxidation and will result in less ROS<sup>142</sup>. This mechanism supports the reduced mutation rates of the *sdhA* evolved alleles,  $\Delta sdhA$ ,  $\Delta sdhB$ , and  $\Delta sdhC$  – these mutations lead to complete inactivation of succinate dehydrogenase, less spurious ROS formation, and reduced mutation rates. However,  $\Delta sdhD$  did not have a low mutation rate (**Figure 3.4B**). In the Sdh protein complex, SdhC anchors SdhAB and is tethered by membrane lipids. Therefore, if *sdhC* is deleted, there is no assembly of SdhAB and similarly, if *sdhB* is absent then SdhA cannot bind to SdhC. The presence of these subunits is interrelated, except with SdhD<sup>122</sup>. It functions to further stabilize the enzyme complex but is not required for activity. Deletion of *sdhD* likely results in an enzyme with partial activity, that remains a producer of ROS. Therefore, the *sdhA* evolved alleles are knockouts of the *sdhA* subunit, that causes inactivation of the whole enzyme, less intracellular ROS, and low mutation rates.

Deletion of this ROS-producing enzyme induces modifications in flux through central carbon metabolism. Primarily, there is a metabolic shift towards the pentose phosphate (PP) pathway during oxidative stress in order to produce more NADPH and less NADH<sup>126,139,143</sup>. Increasing intracellular NADPH concentrations is a key defense against



ROS because it is an essential cofactor for many enzymes that attenuate ROS<sup>144</sup>. Unlike NADH, NADPH will not promote flux through the ETC or the Fenton reaction, which generates OH<sup>-139</sup>. Since there is less flux through the ETC, then general stress response, which is coordinated by *rpoS*, remains uninduced resulting in no mutagenic repair and further explains the low mutation rates of the evolved *sdhA* alleles (**Figure 3.4B**)<sup>145</sup>. In sum, the evolved *sdhA* alleles are deletions of succinate dehydrogenase, which cause less ROS generated by enzyme autoxidation and an upregulation of the PP pathway to produce more NADPH, further combating mutations produced by ROS.

As demonstrated in this discussion of metabolic reprogramming, the *sucD* and *sdhA* antimutator alleles prime the cell's intracellular environment to hamper oxidative stress formation. Less ROS production is achieved by accumulation of metabolic intermediates with antioxidant properties and redirection of central carbon metabolism to concurrently increase the intracellular NADPH:NADH ratio. Ultimately, less ROS translates to less DNA damage and explains the low mutation rates in our PResERV evolved antimutator strains<sup>146</sup>.

### **Rational engineering of antimutators by targeting ROS pathways**

Based on this knowledge, we can begin to answer the following corollary – can we rationally engineer antimutators by reducing intracellular ROS? The NADPH:NADH ratio is central to ROS mitigation. This ratio can be artificially increased by upregulating flux through the main generator of NADPH, the PP pathway. Glucose-6-phosphate dehydrogenase (*zwf*) is the branch point where glycolysis diverts into the PP pathway, overexpression of this enzyme may yield more protective NADPH<sup>126</sup>. In addition, cells

contain many redox enzymes that are known to spuriously autoxidize and produce ROS, such as NADH dehydrogenase II (*ndh*) and xanthine oxidase (*xdh*)<sup>147-149</sup>. Similar to succinate dehydrogenase, inactivation of these enzymes could reduce ROS generation and concomitantly, mutation rates. Arguably, the most direct strategy may be reduction of ROS through upregulation of OxyR and SoxRS repair systems. For example, the OxyR pathway contains an antisense RNA, OxyS, that is synthesized in response to H<sub>2</sub>O<sub>2</sub>. Overexpression of OxyS is known to lower mutation rates in the presence of H<sub>2</sub>O<sub>2</sub><sup>150</sup>. However, overproduction of enzymes that directly detoxify ROS, such as dismutases, may be detrimental and likely depend on the degree of overexpression<sup>151</sup>. It is important to note that some of these modifications may have negative secondary effects on growth and/or metabolism.

Our work provides a platform for future engineering of genetic stability. This study is conceptually the simplest iteration of PResERV, because we employed a model organism harboring a synthetic device with minimal parts. PResERV is a universal method that can be expanded to other organisms and more elaborate synthetic devices. These sorts of experiments will provide a clear picture of genetic stability and how organisms deal with heterologous synthetic devices.

## **MATERIALS AND METHODS**

### **Culture conditions**

*E. coli* was grown in 50 mL Erlenmeyer flasks with incubation at 37°C and 120 rpm orbital shaking over a diameter of 1 inch unless otherwise specified. Cells were grown as 10 mL cultures of Lysogeny Broth (LB) following the Miller recipe (10 g/L tryptone, 5 g/L yeast

extract, and 10 g/L NaCl). Media were supplemented with 50 µg/mL kanamycin (Kan), 20 µg/mL chloramphenicol (Cam), 100 µg/mL rifampicin (Rif), 100 µg/mL carbenicillin (Crb), and 100 µM isopropyl β-d-1-thiogalactopyranoside (IPTG) as indicated. For long-term storage, cultures were frozen at –80°C with glycerol as a cryoprotectant to a final concentration of 13.3% (v/v).

### **Strains and plasmids**

The ancestor strain, BW25113, of the KEIO single gene knockout collection was transformed with the pSKO4 plasmid, which harbors a pBR322 origin of replication (ColE1)<sup>22,55</sup>. This plasmid has been previously edited to remove inactivating terminator sequences<sup>22,118</sup>. The reporter for plasmid mutation rates, pTEM-1.D254tag, encodes a non-functional TEM-1 β-lactamase containing a TAG stop codon in the coding region of the gene<sup>63</sup>. It has a pBR322 origin of replication and encodes the *rop* protein for regulation of plasmid copy number.

### **UV mutagenesis**

To create a UV library of BW25113 + pSKO4, cells were initially mutagenized under a range of conditions (0 to 30,000 µJ/cm<sup>2</sup>, every 2,500 µJ/cm<sup>2</sup>) to generate a kill curve. Three independent cultures of BW25113 + pSKO4 were grown overnight until saturation. Cultures were pelleted as 1 mL volumes and resuspended in an equal amount of saline. For each UV treatment condition, 100 µL cell droplets were pipetted onto an empty petri dish and placed inside a UV crosslinker (UVP CL-1000 Ultraviolet Crosslinker) for exposure. Additionally, a control with zero exposure was necessary to calculate the number of cells

in the starting population. Once irradiated, the droplets were serially diluted (3X by 1:100) and plated to yield countable colonies. The death curve was plotted as a function of the treatment condition (in  $\mu\text{J}/\text{cm}^2$ ) by the number of colonies per treatment. As expected, high UV exposure levels yielded few to zero colonies per plate. We used these data to identify the optimal UV treatment to achieve a ~95-99% death rate. For generating a mutant library of BW25113 + pSKO4, we used an exposure of 27,500  $\mu\text{J}/\text{cm}^2$ .

UV libraries were created by growing five independent cultures of BW25113 + pSKO4. For each replicate, ten 100  $\mu\text{L}$  cell droplets were aliquoted on an empty petri dish and irradiated at the appropriate treatment, as defined previously by the kill curve. The droplets were pooled individually per replicate, pelleted, and resuspended in 10 mL of media for recovery. We estimate that our library size is between  $10^6 - 10^7$  mutants. To ensure that the cells were properly mutagenized, we plated aliquots of the non-mutagenized (overnight cultures) and mutagenized cultures in order to verify that a ~95-99% death rate was achieved. Mutated cells were grown overnight and used as the starting populations for PResERV

## **PResERV**

Five independently UV mutagenized populations of BW25113 + pSKO4 (LB-Crb) were used for PResERV. The fluorescence of each population (GFP excitation = 485, emission = 507) was measured daily. Depending on the percentage of fully fluorescing cells in a given population, cells were either directly transferred (1:1000 dilution) or sorted by FACS (BD FACSAria IIIu) prior to transferring into fresh media. We defined fully fluorescent cells (GFP+) by using the fluorescence intensity of wild-type BW25113 + pSKO4 cells. Therefore, the percentage of fully fluorescing cells was determined by gating the PResERV

populations against the wild-type BW25113 + pSKO4 positive control. If the fluorescence of this gate was ~10-15% of the total population, then the cells were sorted before transferring. If the fluorescence was below this threshold, then the cells were directly transferred without sorting.

For FACS sorting, we used a non-toxic membrane dye, FM 4-64 (ThermoFisher, excitation = 515, emission = 640), in order to visualize cells against any media debris. Cell aliquots of 100  $\mu$ L were stained with 5  $\mu$ g/mL of dye and incubated with shaking for 10 minutes at 37°C. Once stained, cells were pelleted and washed with an equal volume of PBS. Cells were sorted by gating both for fluorescence of the membrane dye in addition to fluorescence of the GFP+ control, described above. In order to achieve proper recovery of this population, we sorted cells at a rate of 1,000 – 2,000 events/second and for a total count of  $10^5$  cells. Cells were immediately transferred into media post sorting. After a series of iterative cell sorts, the PResERV populations were plated to isolate single colonies for characterization and next-generation sequencing (NGS).

### **Isolation of clones with increased GFP stability**

To identify candidate clones for further study, we performed a high-throughput screen to measure the stability of pSKO4 as compared to the BW25113 ancestor (LB-Crb). As a proxy for pSKO4 stability, we utilized the percent fluorescence of the pSKO4 GFP reporter. We picked 70 colonies from each of the five PResERV populations and inoculated them into 96-well plates (1 mL culture volume). Every 96-well plate included three replicates of the positive control, BW25113 + pSKO4, and three replicates of an empty plasmid negative control, BW25113 + pDED376. The plates were grown overnight at 250 r.p.m. We assayed fluorescence of the culture plates by staining 100  $\mu$ L of culture with FM

4-64, performing a 1:100 dilution in PBS, and measuring the fluorescence in a flow cytometer (BD LSRII Fortessa). For interrogating fluorescence, we utilized the same gate parameters as in PResERV and collected 10,000 events per sample. Once the fluorescence values were determined via flow cytometry, the plate cultures were diluted by 1:1000 and grown overnight. Fluorescence was monitored daily for nine days, prior to transferring the cultures into fresh media. We determined the stability of pSKO4 in the PResERV evolved clones by plotting the percent fluorescence by the number of generations (cell doublings). There were two criteria for selecting PResERV clones for further study. Firstly, the evolved clone must have the same starting fluorescence as the positive control and secondly, the evolved clone must maintain long-term fluorescence as compared to the positive control.

### **Mutation rate measurements**

Chromosomal and plasmid mutation rates were determined via Luria-Delbrück fluctuation tests<sup>152</sup>. Strains were grown in LB overnight and diluted by 1:10,000 for one 24-hour cycle of preconditioning in LB media. For measuring plasmid rates, LB media was supplemented with chloramphenicol (LB-Cam) to select for the plasmid containing the reporter gene. Following the preconditioning step, cells for each strain were diluted to a concentration of 1000 cells/200  $\mu$ L. This concentration was utilized to start 20 independent cultures per strain, each with a volume of 200  $\mu$ L. Cultures were grown for 24 hours before plating on non-selective and selective media. For each strain, six cultures were plated on non-selective media and 12 cultures on selective media to obtain counts for the population size and the number of mutants in each culture, respectively. When assaying chromosomal rates, the endogenous gene, *rpoB*, was used as the reporter. Independent cultures were plated on LB for the non-selective condition and LB-Rif for the selective condition. To determine

plasmid rates, we used pTEM-1.D254tag, described above<sup>63</sup>. Independent cultures were plated on LB-Cam for non-selective media and LB-Cam + Crb (500 µg/mL) for selective plates. Cultures plated on non-selective plates were diluted appropriately to yield countable colonies and incubated overnight. Cultures plated on selective media were not diluted and incubated for two days. Mutation rates and adjusted p-values (Bonferroni-corrected) were determined via the maximum likelihood method with SALVADOR 2.3 (R package). For each experiment, pair-wise comparisons were done against the wild-type control mutation rate<sup>153,154</sup>.

### **Genome sequencing**

We performed NGS on three clones from each of the five, independent populations. For isolating genomic DNA, we used the PureLink Genomic DNA Mini Kit (Invitrogen). To fragment DNA, we performed a 20-minute digestion with dsDNA Fragmentase (NEB) to obtain fragments of ~250 bp in length. For preparing NGS libraries, we followed a protocol based on the Kappa LTP Preparation Kit manual KR0453 - v3.13 (Illumina). 2 x 150 paired-end genome sequencing was performed using an Illumina HiSeq at the Genomic Sequencing and Analysis Facility, University of Texas at Austin. Data was analyzed using *breseq* (0.27.2a revision ae9020001ca6)<sup>108</sup>.

### **Absolute plasmid copy number**

Strains harboring pTEM-1.D254tag were grown overnight in LB-Cam. Cultures were diluted the following day by 1:100 and harvested in mid-exponential phase ( $OD_{600} = 0.5$ ). We prepared DNA samples using a PureLink Genomic DNA Mini Kit (Invitrogen). For

calculating absolute copy number, we included genome and plasmid standard curves in all of our qPCR assays. The genomic standard curve template was generated by preparing genomic DNA from wild-type BW25113, lacking pTEM-1.D254tag. The plasmid standard curve template was created by isolating plasmid DNA from wild-type BW25113 + pTEM-1.D254tag using a PureLink Quick Plasmid Miniprep kit (Invitrogen). All samples and standards were normalized to 2 µg/mL for qPCR analysis. Genome and plasmid standards were serially diluted by 10-fold in order to span a range of concentrations between 0.2 ng/µL –  $2.0 \times 10^{-5}$  ng/µL. Samples were diluted by 1:100 for qPCR. We selected *ftsZ* as our target gene for the chromosome and *cat* for targeting the plasmid (~120 bp amplicon length). qPCR reactions were performed using SYBR Green PCR Master Mix (ThermoFisher) in a ViiA 7 Real-Time PCR System (ThermoFisher). The thermal cycling program was as follows: initial denaturation for 10 minutes at 95°C followed by 40 cycles of 95°C for 15 seconds and 54°C for 1 minute, and a final melt curve analysis with a temperature gradient of 0.05°C/second from 54°C to 95°C. The fluorescence signal was measured during the 54°C annealing/extension step. Following the qPCR assay, melt curves for each reaction were analyzed to ensure proper amplification of a single product. Plasmid:genomic DNA ratios were calculated from the average CT values of two technical replicates and using the standard curves as a reference.

### **Growth rate measurements**

For LB growth curves, overnight cultures were diluted by 1:1000 into fresh media and preconditioned for 24 hours in 10 mL LB. After a 24-hour period of preconditioning, cells were diluted to 0.005 OD<sub>600</sub> and transferred to a 96-well plate for OD measurements. OD<sub>600</sub> readings were recorded every 15 minutes using a Tecan Infinite M1000 PRO plate reader.



For minimal media growth curves, colonies were inoculated into 5 mL of LB, grown overnight, diluted by 1:1000, and preconditioned for 24 hours in 10 mL of M9 minimal media. Once preconditioned, growth measurements were conducted by diluting cells to an OD<sub>600</sub> of 0.005 and taking time points every hour using a GeneSys 150 UV-Visible Spectrophotometer (ThermoFisher).

### **Bacteriophage transduction**

P1 transductions were performed by first preparing the donor strain as a P1 lysate. A 5 mL culture of the donor strain was grown overnight and diluted by 1:100 into 5 mL LB supplemented with 0.2% glucose and 5 mM CaCl<sub>2</sub>. Cells were grown at 37°C for ~45 minutes until the culture appeared slightly cloudy. At this point, 100 µL of pre-prepared phage stock was added to the culture. We grew the culture for ~3 hours until lysis was visible and 200 µL of chloroform was added to kill any remaining cells. The lysate was harvested and the supernatant was retained for transduction. To perform a transduction, we grew an overnight culture of the recipient strain, harvested 1.5 mL of cells, and resuspended in 0.75 mL volume with P1 Salts Solution (10 mM CaCl<sub>2</sub>, 5 mM MgSO<sub>4</sub>). 100 µL of recipient cells were mixed with 10 µL of donor lysate and incubated, with no shaking, at 37°C for 30 minutes. Following the incubation, we recovered cells in LB supplemented with 200 µL of sodium citrate for 1 hour, shaking at 37°C. Cells were harvested and the entire mixture was plated on LB-Kan. This protocol is adapted from Thomason *et al*<sup>110</sup>.

### **Strain construction for reversion mutants**

Reversion mutants were made via P1 bacteriophage transduction. We constructed donor strains by genomically linking a *kanR* marker to an evolved allele. Markers were obtained from the i-Deconvoluter library which consists of a series of genomic FRT-Kan-FRT cassettes spaced every ~50 kb in intergenic regions. Given the ultra-dense nature of this library, we were able to insert markers within 50 kb of the mutated site<sup>109</sup>. These strains were used to transduce into the recipient evolved clones, containing the evolved alleles for reversion. Once transduced, recipient strains were sequenced to determine whether the evolved site was reverted back to wild-type. Strains containing the reverted allele were used for mutation rate measurements.

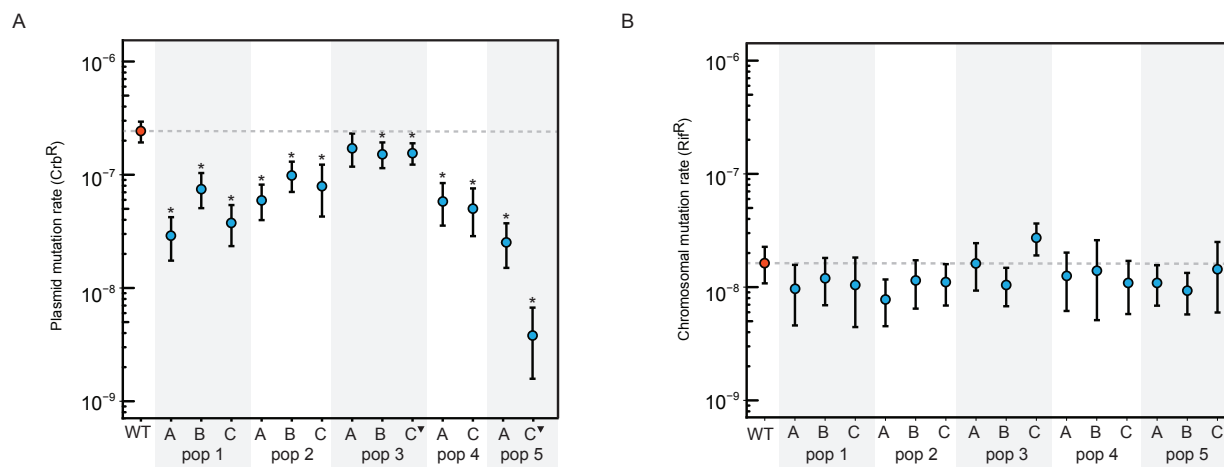
### **Paraquat survival assay**

Overnight cultures in 5 mL LB were started from colonies and diluted by 1:1000 into fresh media. Cells were grown for 12 hours before treatment with 1.5 mM paraquat. Prior to treatment, cultures were diluted in saline and plated on LB to determine forming units (CFUs) before paraquat exposure. Paraquat-treated cultures were grown for another 12 hours and then plated determine CFUs. The ratio of CFUs on treated:untreated plates was calculated per replicate and used to infer survival in paraquat.

### **Intracellular ROS assay**

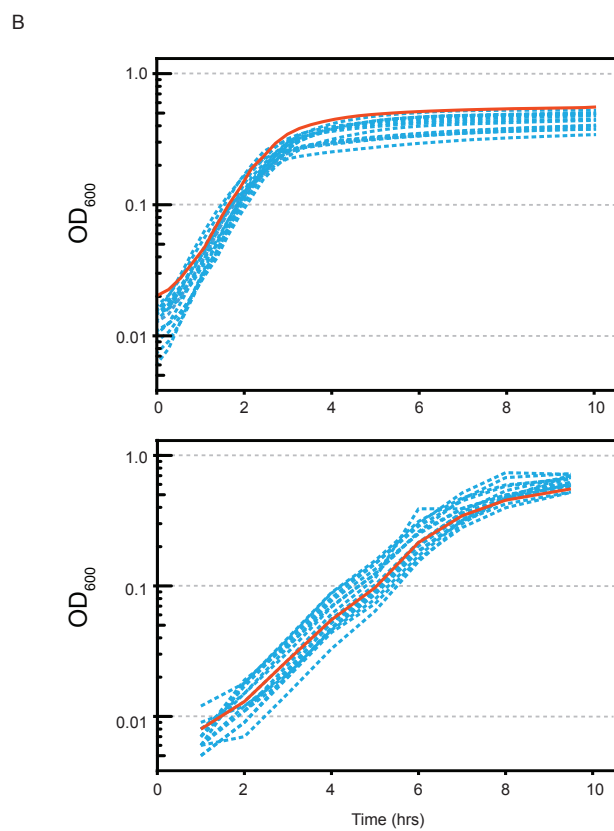
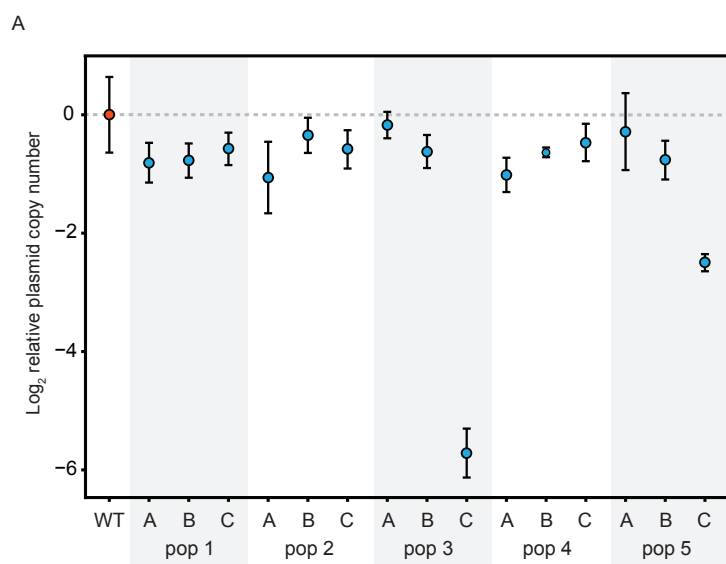
Overnight cultures in 5 mL of LB were diluted by 1:1000 into 50 mL of LB media and grown for 36 hours. ROS measurements were performed by harvesting 500 µL of cells for 5 minutes (3,000 rpm) at 4°C. After this point, all samples were kept on ice. Pellets were washed twice with 500 µL of cold ROS phosphate buffer (137mM NaCl, 2.7mM KCl,

10mM Na<sub>2</sub>HPO<sub>4</sub>, 2mM KH<sub>2</sub>PO<sub>4</sub>, pH = 7.3). Cell pellets were resuspended in 1 mL of ROS buffer and 2 mM of the redox-active dye, 2',7'-Dichlorofluorescein diacetate (Sigma), was added. Samples were incubated in a 37°C heat block for 1 hour. After incubation, samples were subsequently washed with ROS buffer once and fluorescence was measured using flow cytometry (BD LSRII Fortessa, excitation = 490, emission = 519). Data were gated for positive values only, which resulted in a different number of events for each sample. For **Figure 3.5B**, we combined 2500 events at random for each of the four biological replicates and pooled them to plot as one fluorescence distribution. Statistical significance was determined by performing a Mann-Whitney *U* test with this dataset.



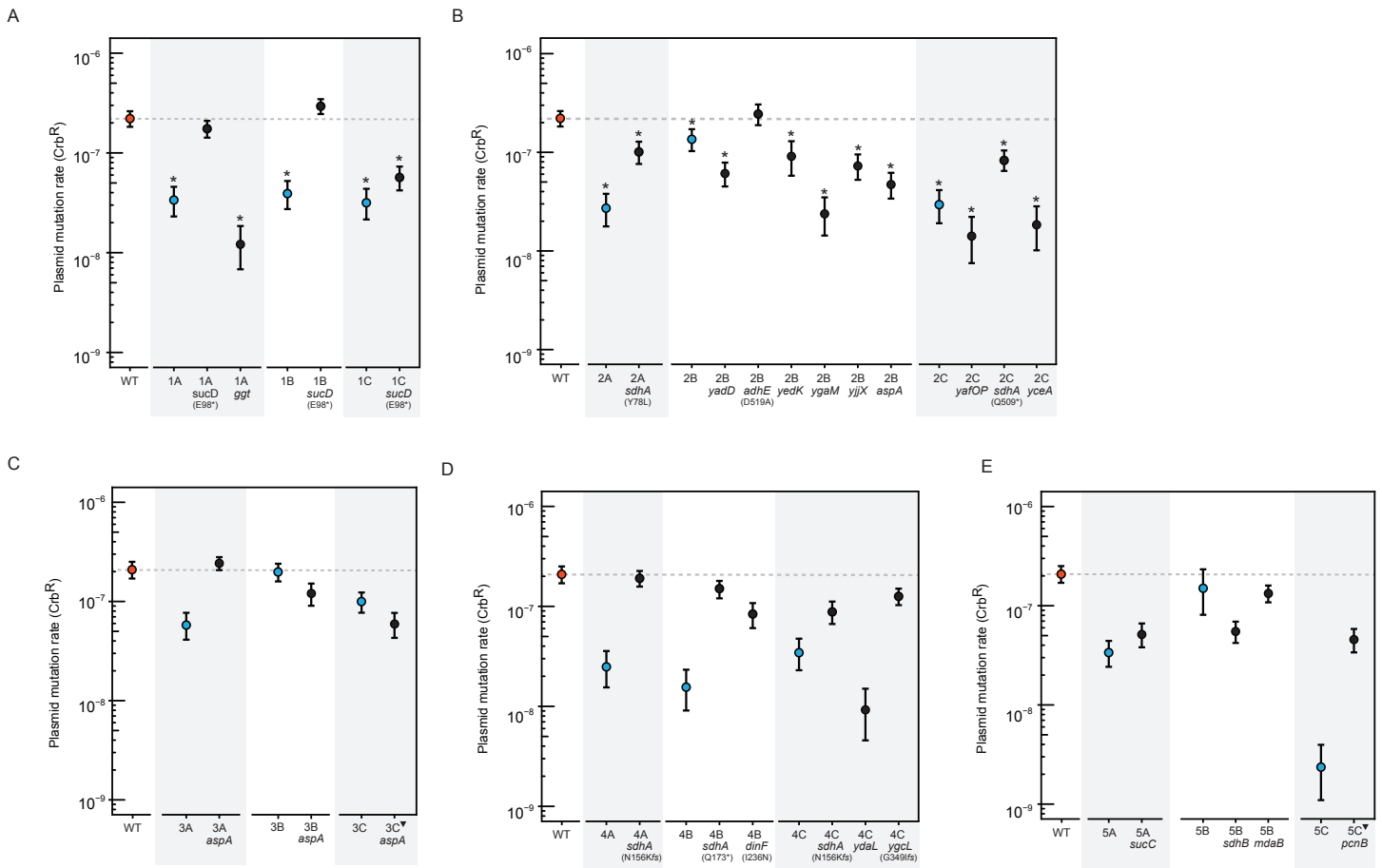
**Figure 3.1. PResERV evolved strains have reduced mutation rates**

(A) Plasmid mutation rates to carbenicillin resistance (Crb<sup>R</sup>) were measured in the evolved PResERV strains via a reporter plasmid containing a nonsense mutation in  $\beta$ -lactamase (pTEM-1.D254tag). The inverted triangle (▼) next to the strain name designates evolved clones in which the plasmid copy is lower than wild-type (see **Figure 3.2A**). (B) Chromosomal rates to rifampicin resistance (Rif<sup>R</sup>) were reported as mutations in the chromosomal gene, *rpoB*. Each plot of mutation rate measurements contains a dashed horizontal line indicating the *E. coli* wild-type mutation rate for comparison. Asterisks (\*) denote statistical significance as determined by adjusted p-values (Bonferroni correction). Pair-wise comparisons were done against the wild-type control mutation rate. Error bars are 95% confidence intervals estimated by the likelihood ratio methods from Luria-Delbrück fluctuation tests (see **Methods**).



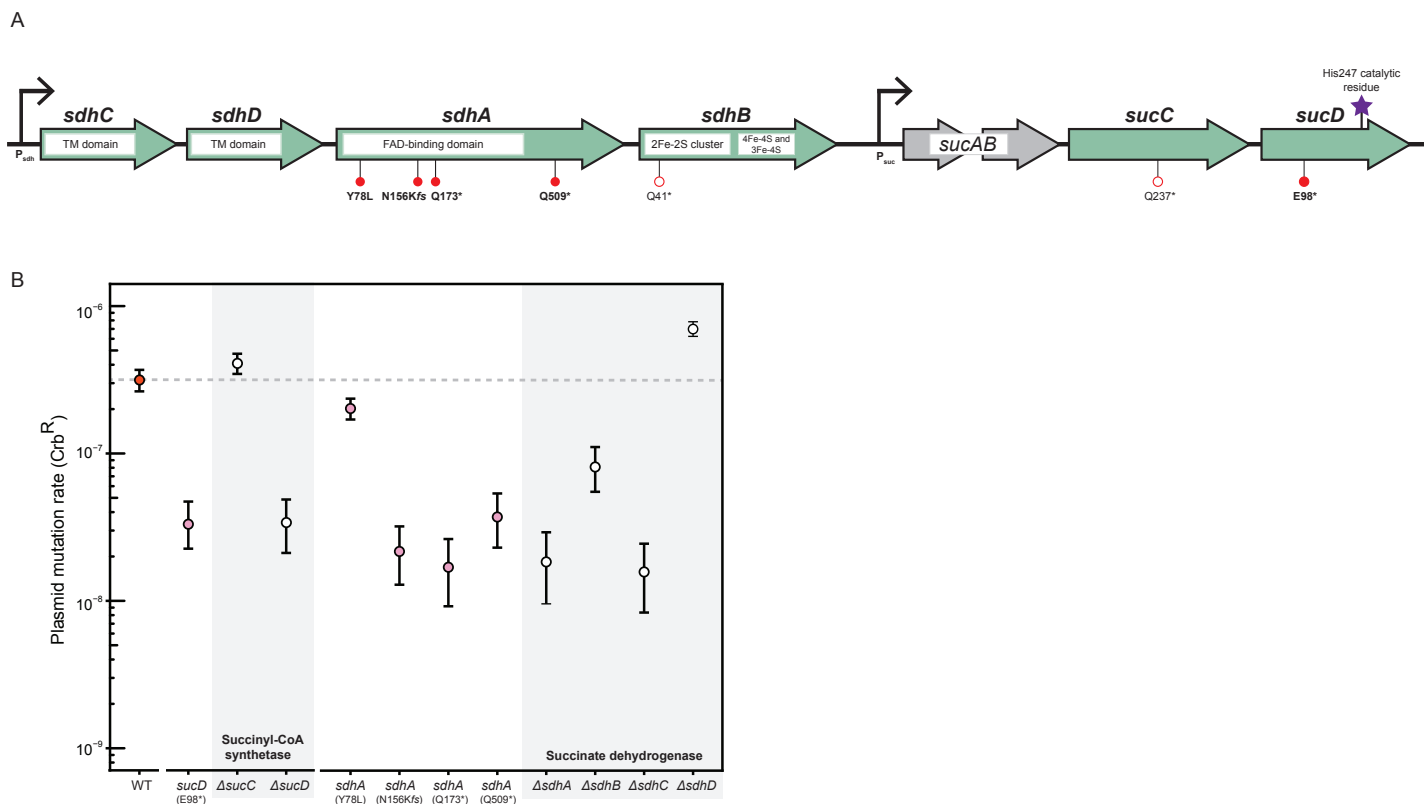
**Figure 3.2. Plasmid copy number and growth rates of PResERV strains**

*Figure 3.2 continued* (A) Plasmid copy number of the evolved strains harboring the mutation rate reporter plasmid (pTEM-1.D254tag) determined by qPCR. The horizontal dashed line indicates the estimated copy number in the wild-type strain. Error bars show the standard error of the mean on log-transformed values from three biological replicates. (B) Growth curves of wild-type (solid, red) and the evolved clones (dashed, blue) in LB media (top panel) and minimal media (bottom panel). Data for LB curves are averages from two technical replicates. Data for minimal media curves are from a single replicate.



### Figure 3.3. Mapping of causative PResERV mutations

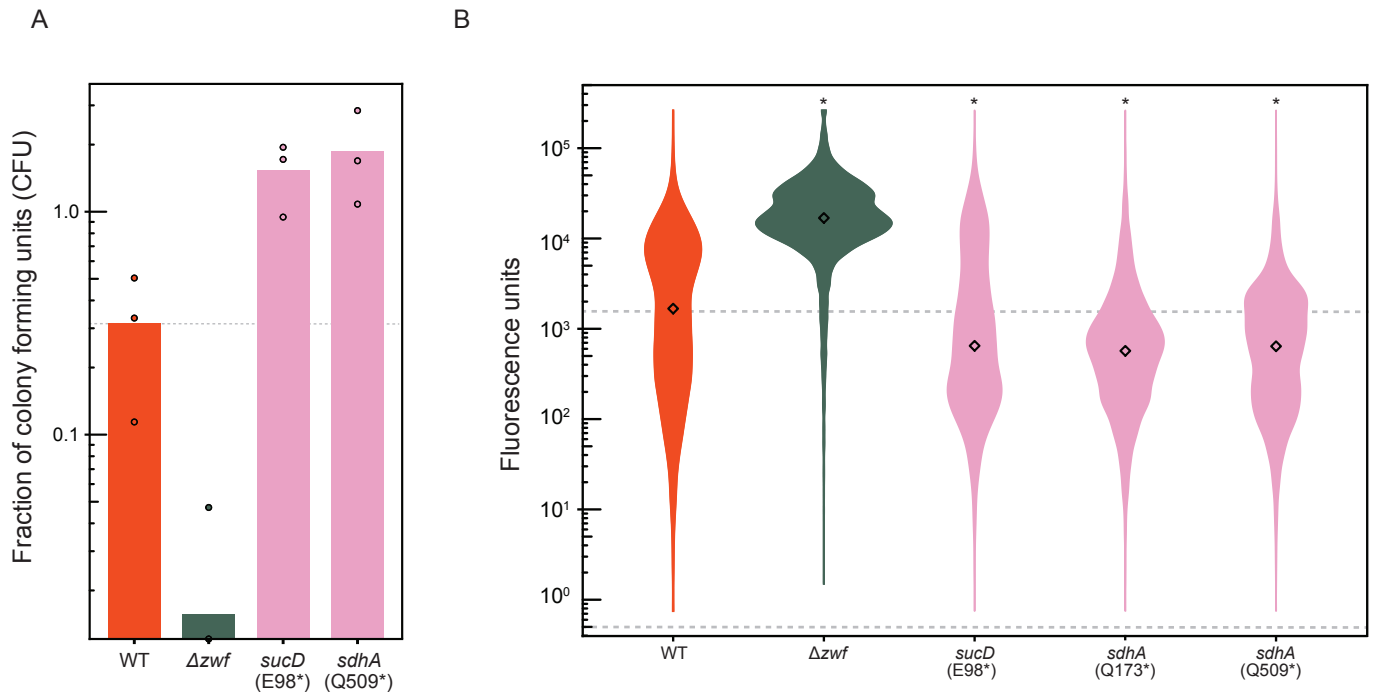
Plasmid mutation rates of wild-type (red and dashed horizontal line), evolved clones (blue), and the respective evolved clone containing a single allele reversion back to wild-type (black). The gene encoding the reverted allele is indicated for each revertant. For causative mutations, the allele information is given. Wild-type measurements are the same data point in (A) and (B) and in (C), (D), and (E). In (C) and (E), a causative mutation could not be identified. The inverted triangle (▼) next to the strain name designates evolved clones in which the plasmid copy is lower than wild-type (see **Figure 3.2A**). Asterisks (\*) denote statistical significance as determined by adjusted p-values (Bonferroni correction). Pair-wise comparisons were done against the wild-type control mutation rate. Error bars are 95% confidence intervals estimated by the likelihood ratio methods from Luria-Delbrück fluctuation tests (see **Methods**).



**Figure 3.4. *sdhA* and *sucD* antimutator alleles**

(A) Operon structure of *sdhCDAB* and *sucCD*. All PResERV mutations, causative (filled red circles) and non-causative (empty red circles), identified in this gene cluster are shown. Functional information is provided for both gene clusters. For the *sdh* genes, the role or cofactors present in each subunit is indicated. For the *suc* genes, the catalytic His247 residue is represented by a star. (B) Plasmid mutation rates of wild type (red and dashed horizontal line), the *sucD* and *sdhA* causative alleles reconstructed in a wild-type background (pink), and the respective deletion mutants of the associated protein complex (white). The specific evolved allele and deletion mutant is indicated. Asterisks (\*) denote statistical significance as determined by adjusted p-values (Bonferroni correction). Pair-wise comparisons were done against the wild-type control mutation rate. Error bars are 95% confidence intervals estimated by the likelihood ratio methods from Luria-Delbrück fluctuation tests (see **Methods**).

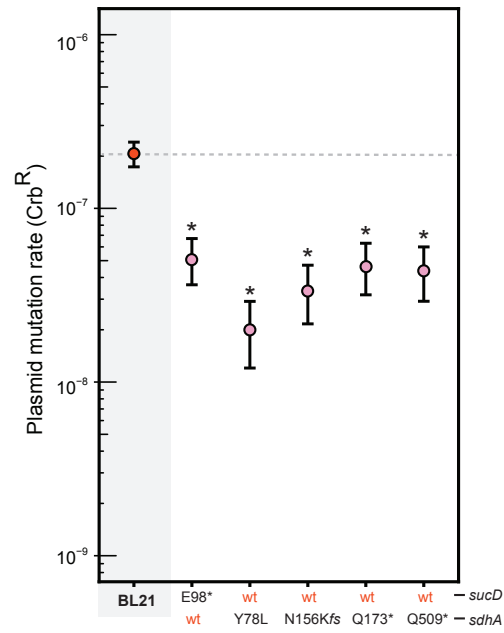




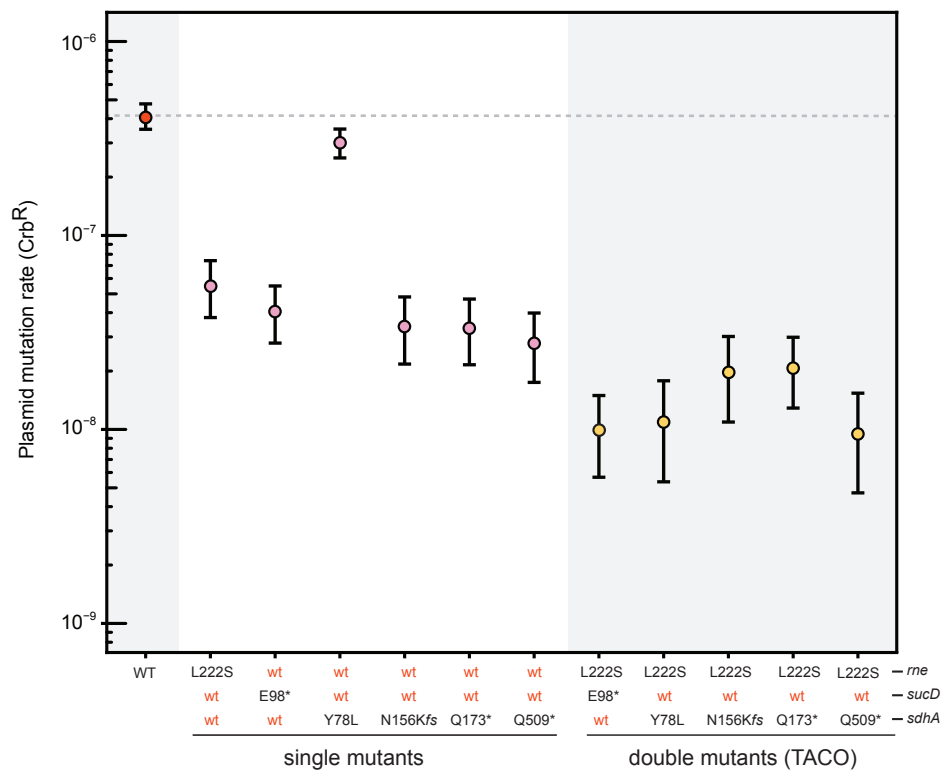
**Figure 3.5. Antimutator *sdhA* and *sucD* alleles function to relieve oxidative stress**

(A) Survival in the presence of a redox-cycling drug, paraquat, of wild type (red and dashed horizontal line),  $\Delta zwf$  (green), and the evolved *sucD*/*sdhA* alleles in wild-type background (pink). CFU counts were determined before and after paraquat treatment. Each circle represents the fraction of surviving cells from a single biological replicate. Bars show the average of the three replicates. (B) Intracellular ROS measurements of wild type (red and top dashed horizontal line indicates median),  $\Delta zwf$  (green), and the *sucD*/*sdhA* alleles in wild-type background (pink). Cells were exposed to a redox active fluorescent dye; high fluorescence values signify elevated levels of intracellular ROS. Data are combined fluorescence values of four biological replicates for each sample and were gated for positive values only, designated by the bottom dashed horizontal line (see **Methods**). Asterisks (\*) denote statistical significance as compared to wild type (Mann-Whitney *U* test). Diamonds within each violin plot represent median values.

A



B



**Figure 3.6. Antimutator alleles reduce mutation rates when ported and combined**

*Figure 3.6 continued* (A) Antimutator *sucD* and *sdhA* alleles were introduced into a commonly utilized *E. coli* protein expression strain, BL21, by transduction. Plasmid mutation rates of wild-type BL21 (red and dashed horizontal line) and BL21 allele exchange mutants containing the reconstructed *sucD* and *sdhA* antimutator alleles (pink). (B) The *sucD* and *sdhA* alleles were combined with a previously identified antimutator allele, *rne*, to build double mutant strains (TACO)<sup>12</sup>. Plasmid mutation rates for wild-type (red and dashed horizontal line), the single antimutator mutants (pink), and the double TACO mutants (yellow) were determined. Asterisks (\*) denote statistical significance as determined by adjusted p-values (Bonferroni correction). Pair-wise comparisons were done against the wild-type control mutation rate. Error bars are 95% confidence intervals estimated by the likelihood ratio methods from Luria-Delbrück fluctuation tests (see **Methods**).

COG Classification	gene	Population 1			Population 2			Population 3			Population 4			Population 5		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Energy production and conversion	<i>sucC</i>													●		
	<i>sucD</i>	●	●	●												
	<i>sdhA</i>				▲		●				■	●	■			
	<i>sdhB</i>										●				●	
	<i>adhE</i>				▲											
Translation, ribosomal structure and biogenesis	<i>yceA</i>						▲									
	<i>pcnB</i>															■
Amino acid transport and metabolism	<i>aspA</i>				▲			●	●	●						
	<i>ggt</i>	▲														
	<i>frtA</i>	▲	▲	▲												
Replication, recombination, and repair	<i>ygcL</i>												■			
	<i>yadD</i>						▲									
Transcription	<i>yafP</i>						■									
	<i>ulaR</i>	◻	◻	◻												
Post-translational modification, protein turnover, and chaperones	<i>btuE</i>						▲									
Defense mechanisms	<i>dinF</i>										▲					
Inorganic ion transport and metabolism	<i>trkD</i>									▲						
Intracellular trafficking, secretion, and vesicular transport	<i>hofC</i>	▲	▲													
Function unknown	<i>mdaB</i>														▲	
	<i>ygaM</i>						▲									
	<i>yjiX</i>						▲									
	<i>yedK</i>						▲									
	<i>ydaL</i>												■			
Total number of mutations		9	6	5	5	18	15	1	1	2	6	6	11	1	6	1

Mutation type	
Base substitution	
●	Synonymous
▲	Nonsynonymous
●	Nonsense
■	Intergenic
Small Indel (<50 bp)	
▲	In-frame
■	Frameshift

**Table 3.1. Candidate antimutator alleles identified via PResERV**

Alleles are grouped by Clusters of Orthologous Groups (COG) function. The legend describes the type of mutation based on a given symbol. Colored symbols illustrate identical mutations found in different strains whereas unique mutations are represented as black symbols.

## Chapter 4: PResERV antimutators evolved from a clean-genome strain, MDS42

### INTRODUCTION

Microbial genomes carry a multitude of mobile genetic elements that are highly disruptive to genetic stability. Movement of these elements creates problems for engineering efforts because they inactivate carefully crafted synthetic devices. For this reason, clean-genome host organism variants have been constructed that are unencumbered by mobile genetic elements<sup>48,50,117</sup>. The *E. coli* version, MDS42, has been deleted of insertion elements (IS) and other seemingly superfluous regions, such as prophage genes, resulting in a ~ 15% reduction in genome size. MDS42 has several interesting properties as compared to its progenitor – similar growth rate, reduced chromosomal mutation rate (1.5-fold), increased transformation efficiency and protein production, and resistance to IS-mediated inactivation of synthetic devices<sup>48,107</sup>. Given these improved properties, we wondered to what extent PResERV would be able to further decrease mutation rates in the MDS42 strain. Therefore, we repeated PResERV on this clean-genome strain in the same way as described in **Chapter 2** and **Chapter 3** using the burdensome pSKO4 plasmid. Using the improved MDS42 host, in which an entire category of inactivating mutations have been eliminated by rational genome engineering, as a starting point for PResERV let us further probe the limits to reducing mutation rates by directed evolution and compare the antimutator alleles found in this host strain to those we found in the wild-type *E. coli* host (BW25113).

## RESULTS

### **Rapid inactivation of a costly plasmid in *E. coli***

Synthetic devices can become non-functional when propagated inside of mutating, living cells. This process depends on the metabolic burden of the synthetic device: the higher the cost of the device, the shorter its lifetime in a lineage of replicating cells. One of the initial steps in PResERV is to characterize how this process happens with the host cell and synthetic device combination of interest. Measuring decay of a particular device is essential in order to know how often transfers should be performed and more importantly, when the population is maintaining device function for longer than the ancestor, at which point putative antimutator clones should be isolated and sequenced.

We measured the decay of the synthetic device used in this PResERV experiment, pSKO4, in the clean-genome strain MDS42 (**Figure 4.1A**). GFP fluorescence from this plasmid was relatively unstable in MDS42. Fluorescence was completely lost in all cells in the population within ~55 generations. For reference, inactivation of pSKO4 in BW25113 occurred after ~75 generations (**Figure 4.1B**). It was surprising that pSKO4 was more stable in BW25113 since MDS42 is purportedly an improved host strain. It has been shown to maintain the integrity of other costly plasmids for longer than its wild-type progenitor strain<sup>48,107</sup>. One possible explanation for our observation with pSKO4, is that MDS42 is designed to guard against IS element-mediated mutations and pSKO4 inactivation may predominantly occur through other mutational mechanisms, such as point mutations. Other synthetic devices that are not often inactivated to mobile element insertions, such as pSKO4, may not have increased stability in MDS42<sup>22</sup>. Performing PResERV directed evolution with MDS42 provides a way to further improve this strain and prevent device inactivation by other mutational mechanisms.

## **PResERV evolution of MDS42**

PResERV was applied to four independent populations of MDS42 + pSKO4, designated MDS pop 1 through 4 (**Figure 4.2**). We set several guidelines for the PResERV evolution experiment. The initial time point shows fluorescence of the population after recovery from UV mutagenesis. Following the first transfer event, the fluorescence of the population drastically decreases in all populations. Possibly, pSKO4 is inactivated in many cells in the UV mutagenized population and these mutants subsequently outcompete those still expressing the plasmid. To enrich for the GFP<sup>+</sup> population, we iteratively sorted for fluorescent cells during each of the subsequent 3-4 transfer events so that they became a majority of the population. Thereafter, regular serial transfers were performed for several days until fluorescence again dropped such that < 10-15% of the population was GFP<sup>+</sup>, at which point a single round of sorting was added to re-enrich for the GFP<sup>+</sup> cells.

In some cases, after a sort and recovery growth period, the population does not reach 100% fluorescence. In part, this may be due quick inactivation of pSKO4 by some mutants prior to them outcompeting the GFP<sup>+</sup> population. It is also a function of how the data is plotted (see **Methods** for details). Fluorescence of a population measured at the single-cell level is a distribution. To summarize these data, we gated each population based on the fluorescence of the positive control, an overnight culture of the ancestral strain-plasmid combination: MDS42 + pSKO4. If the peak of an evolved population is slightly shifted down in fluorescence, only a certain percentage of the population will be counted in the positive control gate. However, this does not mean that the rest of the population is non-fluorescent, as might be assumed from looking at the GFP decay plots. There still exist GFP-expressing cells in these populations. They are just not as highly fluorescing as the positive control. Our criteria for stopping the PResERV evolution experiment was determined by whether the population could maintain fluorescence throughout multiple

transfer events. This improvement versus the ancestral MDS42 + pSKO4 combination was eventually observed for all PResERV populations (**Figure 4.2**).

### **Genetic stability of MDS42 PResERV evolved clones**

As an initial step in characterizing the evolved populations, we isolated individual clones and measured the rate of inactivation of pSKO4. We assayed eleven clones per population (**Figure 4.3**). For each population, we selected clones that maintained fluorescence for longer than the MDS42 ancestor for further characterization. Fluorescence distributions of the positive control, MDS42 + pSKO4, and the negative control, MDS42 + an empty plasmid, are shown in **Figure 4.4**. MDS42 inactivation of pSKO4 is bimodal, however, there is not a complete inactivation of plasmid function, as measured by GFP fluorescence. If the GFP was completely rendered non-functional, its fluorescence distribution would be similar to the empty plasmid control. The data indicate that pSKO4 is reliably mutated to be entirely non-fluorescent in wild-type MDS42 well before 95 generations and therefore we used this as our benchmark for selecting candidate antimutator clones. We selected clones that displayed higher fluorescence distributions at this time point and similar fluorescence values as the positive control at 45 generations (Day 1) (**Figure 4.5**). These evolved clones were further characterized via mutation rate measurements and whole genome sequencing.

### **Mutation rates of evolved clones**

Given that the long-term maintenance of GFP<sup>+</sup> could be a product of reduced mutation rates, we assayed plasmid and chromosomal mutation rates of the 15 MDS42 evolved



clones (**Figure 4.6**)<sup>118</sup>. Thirteen of these evolved clones had modest reductions in plasmid mutation rates ranging from 1.5 to 4.5-fold, and reductions of chromosomal rates ranging from 1.5 to 12-fold. In general, the MDS42 antimutators had greater reductions in the chromosomal mutation rate versus the plasmid. Conversely, the antimutators described in **Chapter 2** and **Chapter 3** had considerable reductions in the plasmid mutation rate and mild reductions in chromosomal rates. The wild-type plasmid mutation rate of MDS42 was ~10-fold lower than BW25113 and the chromosomal rate was ~2-fold lower.

A set of isolates from MDS population 2 exhibited higher plasmid and chromosomal mutation rates. Despite the high mutation rate, these strains could maintain long-term fluorescence of pSKO4. Therefore, they have possibly evolved an alternative method to alleviate the plasmid cost, unrelated to mutation rates. These strains were not studied further given that the goal was to find mutations that increase genetic stability for better maintenance of synthetic devices.

### **Whole genome sequence analysis of MDS42 evolved clones**

To understand the genetic basis of these traits, we sequenced the genomes of the 15 MDS42 evolved strains. **Table 4.1** contains a complete list of all mutations. In the population that evolved hypermutation, MDS population 2, all of the sequenced clones contained a nonsynonymous mutation in *mutT*, a housekeeping enzyme encoding an 8-oxo-dGTP diphosphatase that functions to prevent incorporation of oxidatively damaged nucleotides into DNA/RNA<sup>87</sup>. Mutations that inactivate or reduce *mutT* expression are known to confer a hypermutator phenotype<sup>155,156</sup>. PResERV evolved clones carrying the *mutT* allele contained more mutations than other MDS42 PResERV clones, which makes it difficult to assign the causative mutation. Additionally, too severe bottlenecking of the population may

have aided the *mutT* hypermutator to fix in the population, especially if increased mutation rates helped it evolve to better grow in the medium. Due to this phenotype, mutations in this population were discarded as possible candidates for causative antimutator alleles.

The most attractive antimutator candidates were found in MDS populations 1 and 4. MDS population 4 strains (DL712, DL714, DL698, DL701, DL703, and DL707) carry a nonsense mutation in *mdh*, which encodes for a TCA cycle enzyme, malate dehydrogenase. Our previous knowledge on PResERV antimutator strains and reduction in oxidative stress via TCA cycle enzymes aided in identifying this mutation as a candidate (**Chapter 3**). Malate dehydrogenase catalyzes the oxidation of malate to oxaloacetate and concomitantly generates NADH<sup>157</sup>. NADH supplies the electron transport chain, and therefore a nonsense mutation may result in complete inactivation of the enzyme and reduced oxidative stress. In addition, Mdh is known to facilitate induction of stress induced mutagenesis via *rpoS*, deletion of this enzyme suppresses this pathway and leads to reduced mutation rates<sup>145</sup>. These mechanisms are similar to those discussed for the *sucD* and *sdhA* antimutator alleles in **Chapter 3**. It is likely that cells have adapted many ways to reduced oxidative stress, and therefore *mdh* may be an interesting candidate mutation to further support the involvement of this pathway in genetic stability.

The remaining candidate mutations serve roles in a seemingly more direct connection to genetic stability, DNA/RNA metabolism. Strains from MDS populations 1 and 4 contain nonsynonymous mutations in subunits of the main replicative polymerase in *E. coli* (DNA polymerase III), *dnaX* (DL660, DL662, and DL666) and *dnaE* (DL698, DL703, DL707), respectively. Notably, **Chapter 2** describes causative antimutator alleles found in other polymerases, *polA* and *polB*. There are many examples of known antimutator DNA polymerases, therefore it is not surprising if the MDS42 PResERV *dnaX* and *dnaE* mutations lead to an antimutator Pol III<sup>36</sup>. Antimutators of *dnaE* have been

previously reported (**Chapter 1**)<sup>35</sup>. PResERV evolved clones DL660 (containing the evolved *dnaX* allele) and DL698 (containing the evolved *dnaE* allele) have lower chromosomal mutation rates. However, other clones containing these mutations did not (DL666, DL703, and DL707). These clones may contain other mutations that suppress the antimutator effects of *dnaE* and *dnaX*. Additionally, there is one PResERV clone that remains to be assayed (DL662). Regarding RNA metabolism, strains from MDS population 4 harbor mutations in *rplE* and *deaD* (DL712, DL714, DL698, DL701, DL703, and DL707). RplE is a 50S ribosomal subunit protein that forms the only protein-protein connection between both ribosomal subunits<sup>158</sup>. DeaD encodes an RNA helicase that, similar to the *rne* causative gene discussed in **Chapter 2**, is involved in a myriad of RNA processing activities. In sum, there are interesting candidate mutations found in the MDS42 PResERV antimutators and many of these potentially overlap with antimutator mechanisms illustrated in **Chapter 2** and **Chapter 3**.

## **FUTURE DIRECTIONS**

A key immediate step in this work is to validate the putative causative antimutator mutations by either reverting interesting alleles (such as the ones discussed above) or porting alleles into a wild type background and measuring the effect on plasmid/chromosomal mutation rates. Once identified, specific assays can be performed to discern the mechanism of action. These mutations, along with those identified in **Chapter 2** and **Chapter 3**, can be combined to form different/better versions of TACO and introduced into other organisms to determine whether they still function to reduce mutation rates. Furthermore, there is an even more improved version of MDS42, known as MDS42 LowMut, in which several SOS-induced DNA polymerases have been deleted (Pol II,

PollV, and PolV) leading to a ~3-fold reduction in chromosomal mutation rates<sup>39</sup>. It would be interesting to use this strain for PResERV and/or introduce antimutator alleles from **Chapter 2** and **Chapter 3** into MDS42 LowMut to create a TACO strain with marked reductions in both plasmid and chromosomal mutation rates.

## **MATERIALS AND METHODS**

### **Culture conditions**

*E. coli* was grown as 10 mL cultures in 50 mL Erlenmeyer flasks with shaking incubation at 37°C and 120 r.p.m. unless otherwise specified. Lysogeny Broth (LB, Miller) was used (10 g/L tryptone, 5 g/L yeast extract, and 10 g/L NaCl). Media were supplemented with 100 µg/mL carbenicillin (Crb), 20 µg/mL chloramphenicol (Cam), 100 µg/mL rifampicin (Rif), and 1 mM isopropyl β-d-1-thiogalactopyranoside (IPTG), as indicated. Bacterial cultures were frozen at –80°C with glycerol (13.3% v/v final concentration) as a cryoprotectant for long-term storage.

### **Strains and plasmids**

MDS42, a derivative of *E. coli* K-12 strain MG1655, was transformed with pSKO4. This plasmid contains a pBR322 origin of replication (ColE1) and has been previously reengineered to remove unstable sequences<sup>22</sup>. The mutation rate reporter plasmid, pTEM-1.D254tag, harbors a β-lactamase gene with a nonsense (TAG) stop codon in the coding region, rendering the gene nonfunctional. This plasmid contains a pBR322 origin of replication along with the regulatory *rop* protein<sup>63</sup>.

## **UV mutagenesis**

To create a library of mutants for PResERV, MDS42 cells containing pSKO4 were UV mutagenized. Initially, a pilot experiment was performed under a range of conditions (0 to 30,000  $\mu\text{J}/\text{cm}^2$ , every 2,500  $\mu\text{J}/\text{cm}^2$ ) in order to generate a kill curve and identify the point in which a 95-99% death rate is achieved. Three independent cultures of MDS42 + pSKO4 were grown overnight, cultures were harvested, and the pellets were resuspended in an equal volume of saline. For each UV treatment, 100  $\mu\text{L}$  cell droplets were aliquoted onto an empty petri dish and exposed inside a UV crosslinker (CL-1000 Ultraviolet Crosslinker, UVP). A zero-exposure control was used to calculate the initial cell number. Cells were plated (LB-Crb) to determine colony counts and the kill curve was plotted against the treatment condition (in  $\mu\text{J}/\text{cm}^2$ ). These data were used to identify the optimal UV treatment. For creating a mutant library of MDS42 + pSKO4, we used an exposure of 15,000  $\mu\text{J}/\text{cm}^2$ .

UV libraries were created by growing four separate cultures of MDS42 + pSKO4 from independent colonies. Cultures were grown overnight, 1 mL of each culture was pelleted and resuspended in an equal volume of saline. Per replicate, ten 100  $\mu\text{L}$  cell droplets were pipetted onto an empty petri dish, irradiated at 15,000  $\mu\text{J}/\text{cm}^2$ , and pooled in 10 mL of LB-Crb medium. These libraries were used as the starting populations for PResERV. To verify the ~95-99% death rate, we plated non-mutagenized and mutagenized cells. We estimate that each library contained between  $10^6 - 10^7$  mutagenized cells.

## **PResERV directed evolution**

Four independently UV mutagenized populations of MDS42 + pSKO4 were used as the starting populations for PResERV. Once mutagenized, the library was recovered in fresh media (LB-Crb-IPTG) overnight before beginning transfers. GFP fluorescence (485 nm

excitation, 507 nm emission) was monitored daily via flow cytometry (BD LSRII Fortessa). The fluorescence of an evolved population was compared to the positive control, an overnight culture of non-mutagenized MDS42 + pSKO4. The entire fluorescence distribution of this control was used to gate the evolved populations. For a given population, if the fraction of fully fluorescent cells in this gate was below ~10-15%, then the population was sorted by FACS (BD, FACS Aria IIIu) and the GFP+ cells in this sort gate were transferred into fresh media for regrowth. If the fraction of fluorescent cells in this gate was above the ~10-15% cutoff, cells were transferred without sorting.

To visualize cells against media debris during FACS sorting and flow cytometry, we used a non-toxic membrane dye FM 4-64 (ThermoFisher 515 nm excitation, 640 nm emission). To dye cells, 5 µg/mL of FM 4-64 dye was added to 100 µL of cells and incubated with shaking for 10 minutes at 37°C. Cells were subsequently pelleted and washed with an equal volume of PBS. Cells were sorted by gating for GFP+ fluorescence, as described earlier, and for fluorescence of the membrane dye. Cells were sorted at a rate of 1,000 – 2,000 events/second for a total of 10<sup>5</sup> cells. After sorting, cells were recovered in fresh LB-Crb-IPTG media. Once it was determined that the population was maintaining a population of GFP+ cells for longer than the MDS42 ancestor, cells were plated to isolate single clones for further characterization and sequencing. All flow data (from cytometry and FACS) was analyzed in FlowJo (v10.5.0).

### **GFP decay curves**

Stability of pSKO4 in MDS42 and the evolved PResERV clones was quantified using a GFP decay assay. For each strain, colonies were used to inoculate 10 mL of LB-Crb-IPTG and grown overnight until saturation. Each population was transferred daily under the same

conditions as the PResERV experiment and GFP+ fluorescence was measured prior to each transfer. Cells were stained with FM 4-64 as described above and fluorescence was measured by flow cytometry. For visualizing fluorescence data of the percentage of GFP+ cells in a population over time (as in **Figures 4.1 – 4.3**), cells were classified as GFP+ if they were similar to the MDS42 wild-type distribution, as explained above and this percentage of the population is plotted on the Y-axis. Additionally, the GFP+ cells also had to have wild-type levels of fluorescence when stained with the membrane dye FM 4-64. Therefore, our criteria gates for cells that are highly fluorescent and against any debris. GFP decay curve experiments were used to screen clones after PResERV for follow-up studies and NGS sequencing.

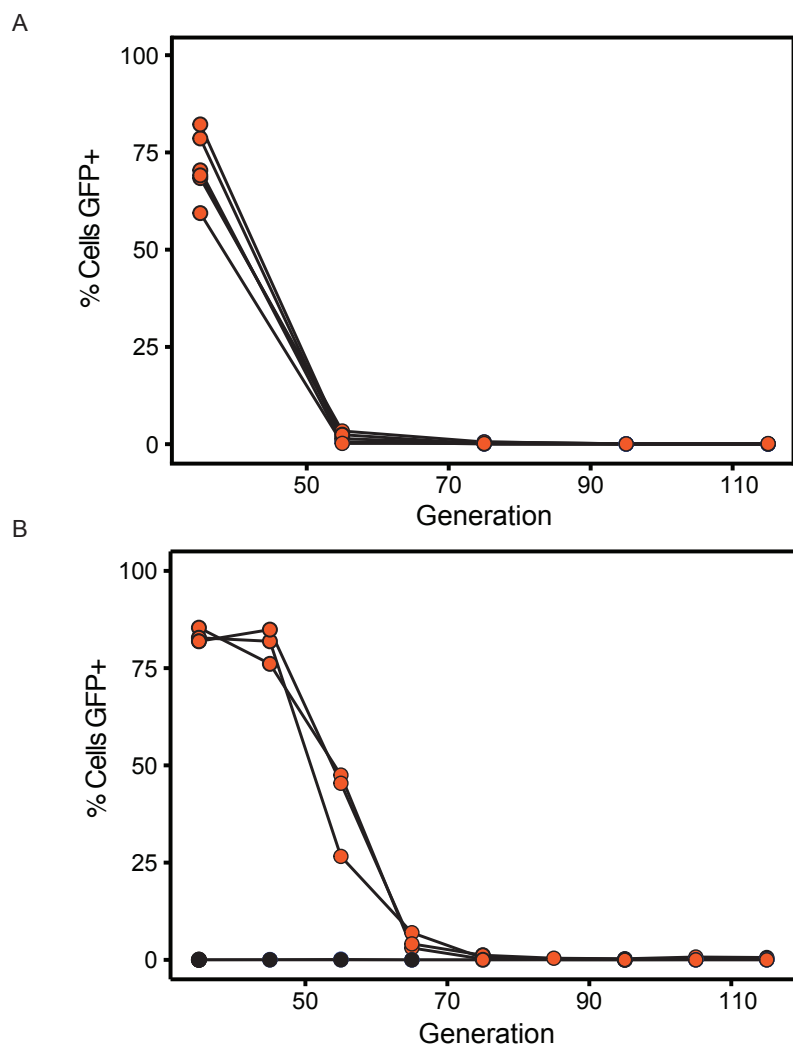
### **Genome sequencing**

We performed next-generation sequencing (NGS) on individually isolated clones from each of the four, evolved populations. DNA was isolated using a PureLink Genomic DNA Mini Kit (Invitrogen). DNA was fragmented by performing a 20-minute digestion with dsDNA Fragmentase (NEB) to generate ~250-bp fragments. NGS libraries were prepared following a protocol based on the Kappa LTP Preparation Kit manual KR0453 - v3.13 (Illumina). An Illumina HiSeq was used to sequence 2 x 150 paired-end reads per sample at the Genomic Sequencing and Analysis Facility, University of Texas at Austin. Data was analyzed using *breseq* (0.27.2a revision ae9020001ca6)<sup>108</sup>. The MDS42 reference genome was obtained from NCBI (NC\_020518.1).

### **Mutation rate measurements**

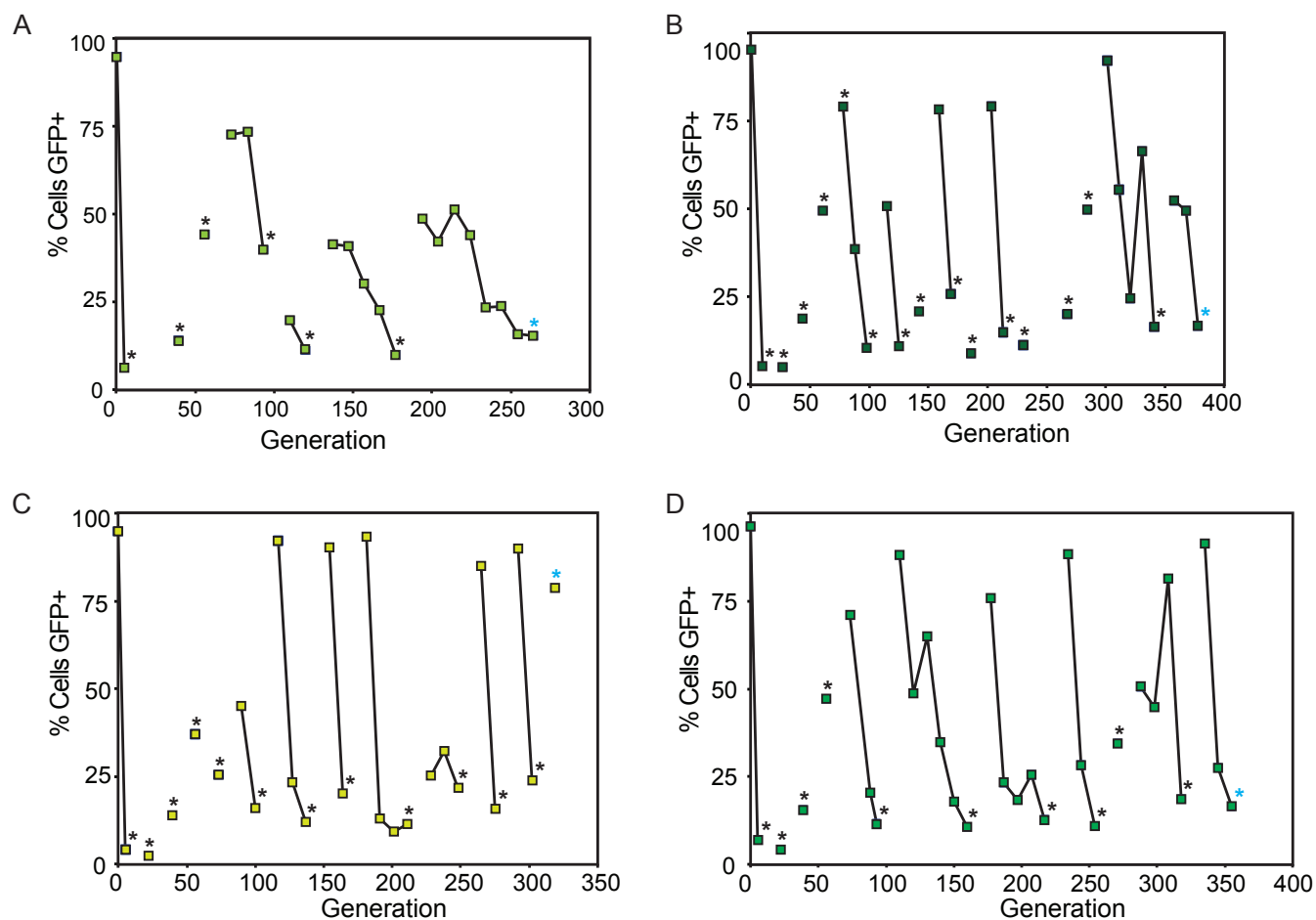
Chromosomal and plasmid mutation rates were measured using Luria-Delbrück fluctuation tests<sup>152</sup>. Strains were preconditioned in 5 mL media overnight and then diluted by 1:10,000 into fresh media and allowed to grow to saturation over an additional 24-hour growth cycle. For measuring plasmid rates, LB media was supplemented with chloramphenicol (LB-Cam) to select for retention of the mutation rate reporter plasmid, pTEM-1.D254tag. After preconditioning in LB, cells were diluted to a concentration of 1000 cells/200  $\mu$ L and used to start 20 independent cultures. Following 24 hours of growth, six cultures were plated on non-selective agar and 12 cultures were plated on selective agar. For chromosomal rate assays, the non-selective medium was LB and the selective medium was LB-Rif. For plasmid rates, the non-selective media condition was LB-Cam and the selective media condition was LB-Cam-Crb (500  $\mu$ g/mL). Cultures that were plated on non-selective plates were diluted in order to yield countable colonies and the agar plates incubated overnight before counting colonies. Cultures that were plated on selective media were not diluted before plating and the agar plates were incubated for two days before counting colonies. Mutation rates were estimated from these colony counts via the maximum likelihood method with rSalvador 2.3<sup>115</sup>.





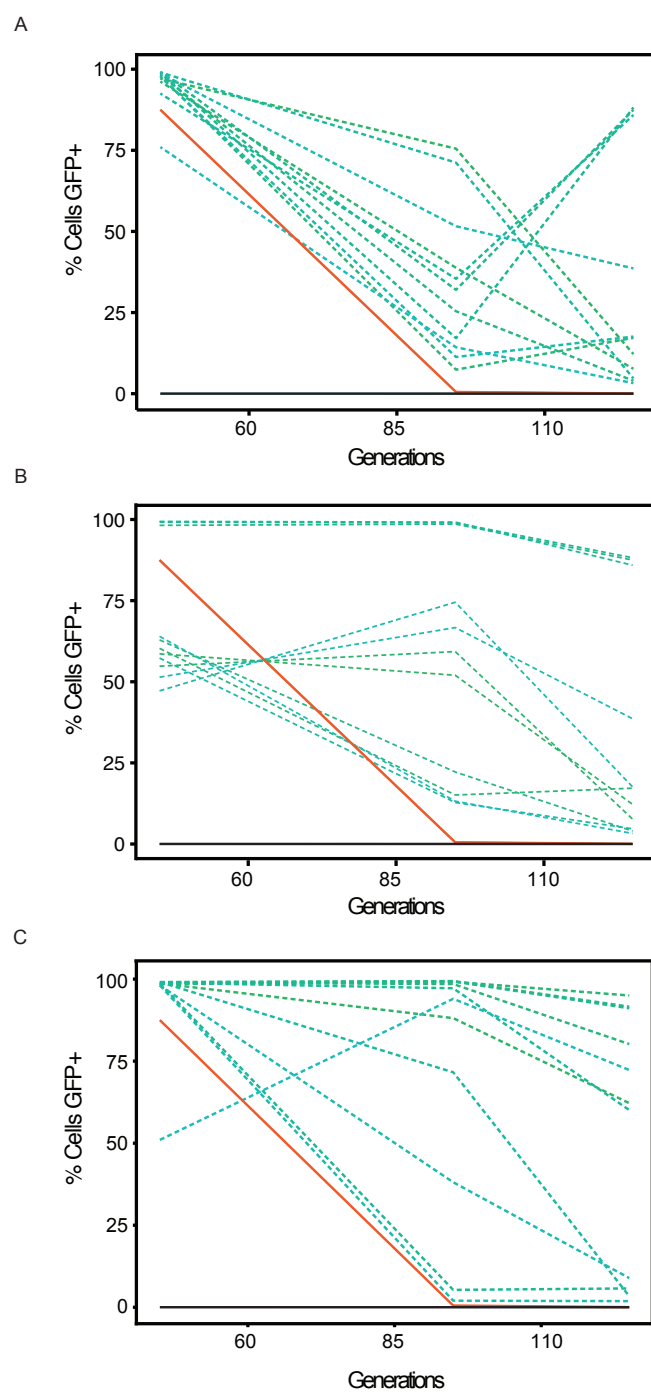
**Figure 4.1. Inactivation of pSKO4 in *E. coli***

GFP decay measurements of (A) five independent MDS42 cultures and (B) three independent cultures of BW25113 (orange). For reference, an empty plasmid control is shown in B (black). Data are plotted as described in **Methods**.



**Figure 4.2. FACS sorting for MDS42 PResERV evolved populations**

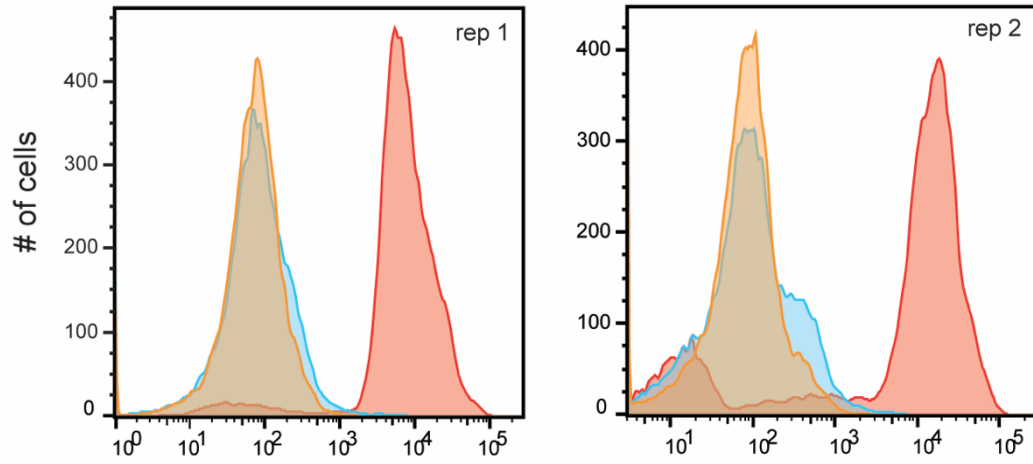
GFP fluorescence during PResERV evolution of (A) MDS42 population 1, (B) MDS42 population 2, (C) MDS42 population 3, and (D) MDS42 population 4. Sort events are denoted by an asterisk (\*). Transfers are indicated by a line connecting each fluorescence measurement. Final plating of the population to isolate clones is designated by blue asterisk (\*). Data are plotted as described in **Methods**.



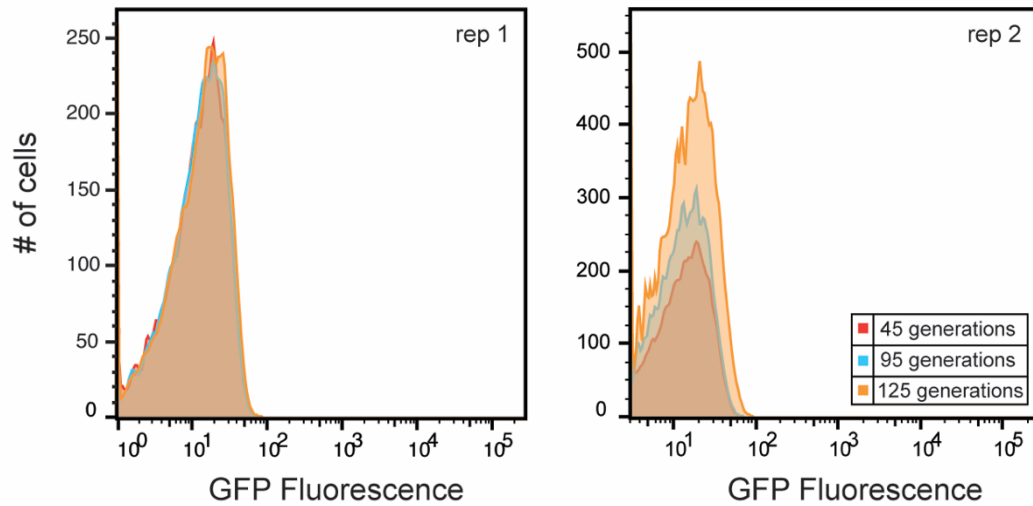
**Figure 4.3. GFP Stability of PResERV evolved isolates**

*Figure 4.3 continued* Individual clones were isolated following PResERV evolution. Decay of pSKO4 is shown for (A) MDS42 population 2, (B) MDS42 population 3, and (C) MDS42 population 4. The solid orange line represents the positive control, MDS42 containing pSKO4, the solid black line represents the negative control, MDS42 containing an empty plasmid, and the dotted lines represent an evolved PResERV clone. Data are plotted as a function of days (transfers) as described in **Methods**.

A

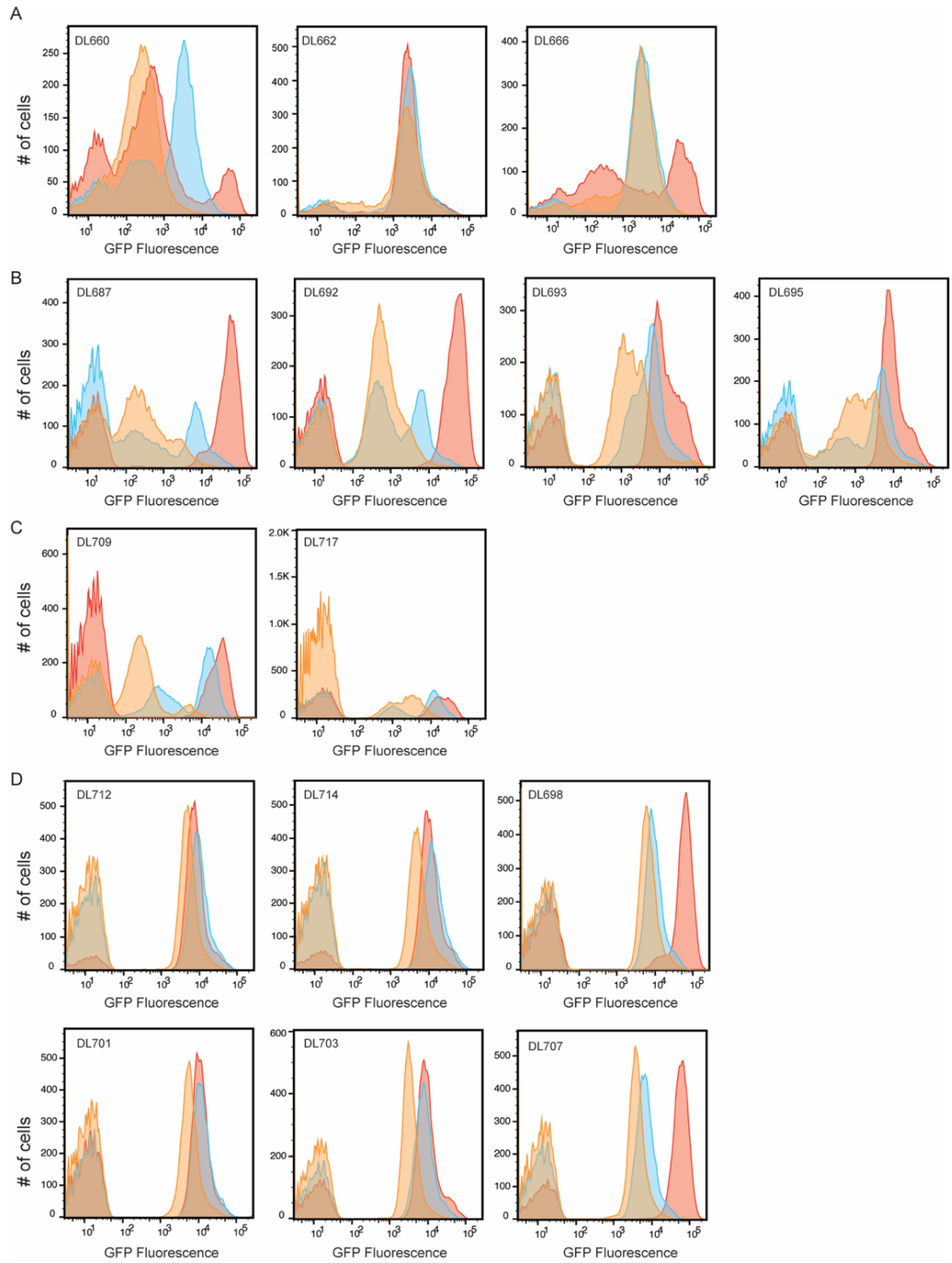


B



**Figure 4.4. Fluorescence distributions of MDS42 controls during a GFP decay experiment**

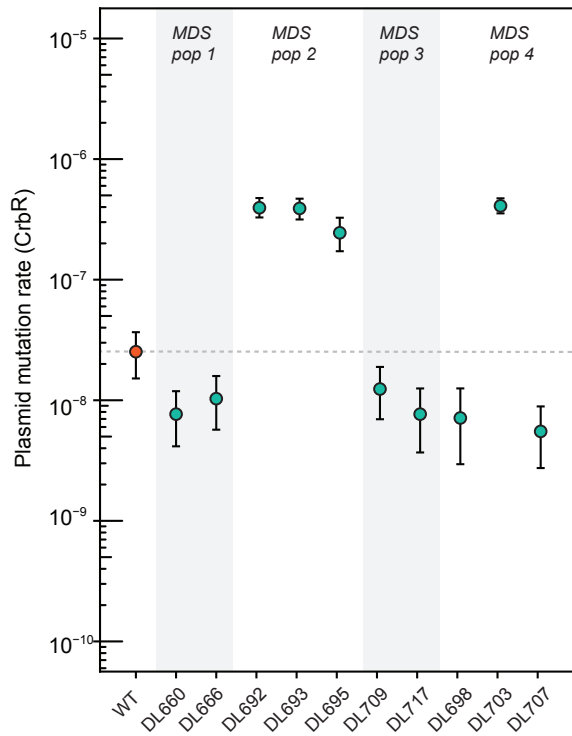
Two biological replicates of the (A) positive control, MDS42 + pSKO4, and the (B) negative control, MDS42 + an empty plasmid. Fluorescence was measured prior to transfer of cells into fresh media. Distributions for 45 generations (red), 95 generations (blue), and 125 generations (orange) are shown.



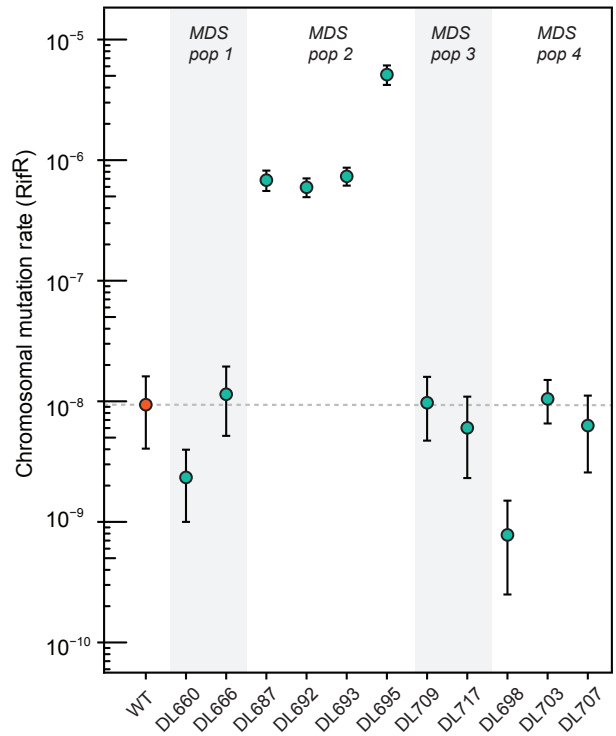
**Figure 4.5. Fluorescence distributions of candidate antimutator strains**

*Figure 4.5 continued* One biological replicate for each selected clone is pictured. Candidate clones from (A) MDS42 population 1, (B) MDS42 population 2, (C) MDS42 population 3, and (D) MDS42 population 4 are shown. Fluorescence was measured prior to transfer of cells into fresh media. Distributions for 45 generations (red), 95 generations (blue), and 125 generations (orange) are shown.

A



B



**Figure 4.6. Mutation rates of MDS42 PResERV evolved clones**

(A) Plasmid and (B) chromosomal mutation rates are shown for the MDS42 evolved clones. Each plot contains a dashed horizontal line indicating the MDS42 wild type (orange) mutation rate for comparison. The evolved clones (blue) are grouped by population. Error bars are 95% confidence intervals estimated by the likelihood ratio methods from Luria-Delbrück fluctuation tests (see **Methods**).



MDS pop 1		MDS pop 2			MDS pop 3		MDS pop 4					annotation	gene	description
							DL712	DL714	DL698	DL701	DL703	DL707	<i>thrA</i> →	aspartokinase/homoserine dehydrogenase
		DL687	DL692	DL693	DL695								<i>mutT</i> →	8-oxo-dGTP diphosphatase
								DL698			DL703	DL707	<i>dnaE</i> →	DNA polymerase III alpha subunit
													<i>dnaX</i> →	DNA polymerase III, tau and gamma subunits
DL660	DL662	DL666											<i>tdra</i> → / → <i>yibF</i>	predicted acyl-CoA synthetase NAD(P)-binding/conserved protein
			DL692										<i>ahpF</i> →	Alkyl hydroperoxide reductase, FAD/NAD(P)-binding
		DL687	DL692	DL693	DL695								<i>citC</i> ← / → <i>citA</i>	Citrate lyase synthetase/sensory histidine kinase with CitB
													<i>ybdN</i> →	Predicted lipoprotein
			DL693										<i>yjgS</i> ← / → <i>arog</i>	Conserved protein/aldolase
				DL693									<i>poxB</i> ←	Pyruvate dehydrogenase, thiamin-dependent, FAD-binding
	DL662				DL695								<i>focA</i> ←	Formate transporter
													<i>mdoG</i> → / → <i>yocK</i>	Glucan biosynthesis protein/predicted lipoprotein
													<i>lpxL</i> ←	ACP-dependent acyltransferase
			DL692										<i>fabH</i> →	ACP synthase III
		DL687	DL692	DL693	DL695								<i>poIA</i> ← / → <i>pepT</i>	Polyamine transporter subunit/peptidase T
					DL695									

Table 4.1. Mutations in MDS42 PResERV strains

MDS pop 1		MDS pop 2				MDS pop 3		MDS pop 4						annotation	gene	description
DL660	DL666													<b>O223H</b> (CA <b>G</b> →CA <b>T</b> )	<i>nhaB</i> ←	Sodium/proton antiporter
		DL687												<b>V156G</b> (G <b>T</b> A→G <b>G</b> A)	<i>treA</i> ←	Periplasmic trehalase
				DL693										<b>D319E</b> (GA <b>T</b> →GA <b>G</b> )	<i>prfA</i> →	Peptide chain release factor RF-1
														<b>M83I</b> (AT <b>G</b> →AT <b>I</b> )	<i>chxA</i> ←	Calcium/sodium:proton antiporter
DL662														<b>P231P</b> (CC <b>A</b> →CC <b>C</b> )	<i>dis</i> ←	Cardiolipin synthase
		DL687												intergenic (-108/+595)	<i>yddV</i> ← / ← <i>yddW</i>	predicted diguanilate cyclase/predicted lipoprotein
														intergenic (-178/+184)	<i>gadB</i> ← / ← <i>pqQL</i>	Glutamate decarboxylase/predicted peptidase
								DL712						Δ1,174 bp	<i>ynfA-[speG]</i>	<i>ynfA</i> , <i>ynfB</i> , <i>[speG]</i>
								DL712	DL714					<b>P29P</b> (CC <b>A</b> →CC <b>G</b> )	<i>nbc</i> ←	Rhino flavin synthase, alpha subunit
										DL698	DL701	DL703	DL707	<b>L286*</b> (TT <b>A</b> →T <b>G</b> A)	<i>torY</i> ←	TMAO reductase III, cytochrome c-type subunit
		DL687	DL692	DL693	DL695									<b>L131R</b> (CT <b>G</b> →C <b>G</b> G)	<i>shfA</i> →	Shikimate transporter
					DL695									<b>A114A</b> (GC <b>T</b> →GC <b>G</b> )	<i>yeoY</i> ←	Predicted DNA-binding transcriptional regulator
														<b>F35C</b> (TT <b>T</b> →T <b>G</b> T)	<i>yegU</i> →	Predicted hydrolase
				DL693										<b>Y115D</b> (T <b>A</b> C→ <b>G</b> A <b>C</b> )	<i>mtaA</i> →	ribonucleoside diphosphate reductase 1, alpha subunit
		DL687	DL692	DL693	DL695									Coding (321/1218 nt)	<i>yfbQ</i> →	Predicted aminotransferase
						DL709	DL717									

Table 4.1 continued

MDS pop 1			MDS pop 2			MDS pop 3			MDS pop 4					annotation	gene	description
					DL693									A76A (GCA→GCQ)	ypfE ←	predicted carboxysome structural protein with predicted role in ethanol utilization
DL660														Intergenic (-46/+126)	gluW ← / ← rrsG	rRNA-Glu/16S ribosomal RNA
DL660														noncoding (273/1542 nt)	rrsG ←	16S ribosomal RNA
														V489V (GTI→GTG)	recN →	Recombination and repair protein
			DL692											V268V (GTI→GTG)	proW →	Glycine betaine transporter subunit
DL660	DL662	DL666												*374C (TGA→TGG)	hnpD →	Protein required for maturation of hydrogenases
			DL692											T319T (ACG→ACI)	amcC ←	N-acetyl/muramoyl-L-alanine amidase
						DL709	DL717							M1M (ATG→CTG)	mscS ←	Mechanosensitive channel
			DL687											L112* (TTA→TGA)	yggH ←	rRNA methyltransferase, SAM-dependent
									DL701					I106L (ATT→CTT)	yglE →	Predicted tartrate:succinate antiporter
			DL687											intergenic (-43/-245)	yglH ← / → yglI	predicted siderophore interacting protein/predicted transcriptional regulator
	DL662													K22Q (AAA→CAA)	yqjG →	Predicted S-transferase
			DL687	DL692	DL693	DL695								coding (1877-1878/1890 nt)	dead ←	ATP-dependent RNA helicase
									DL712	DL714	DL698	DL701	DL703			
			DL687	DL692	DL693	DL695								V446G (GIG→GGG)	nanT ←	Sialic acid transporter
												DL701		intergenic (-74/-306)	nanR ← / → dcuD	DNA-binding transcriptional regulator/ Predicted transporter

Table 4.1 continued

MDS pop 1		MDS pop 2			MDS pop 3		MDS pop 4					annotation	gene	description
							DL712	DL714	DL698	DL701	DL703	DL707	<i>mdh</i> ←	Malate dehydrogenase, NAD(P)H-binding
		DL687	DL692	DL693	DL695								<i>mdh</i> ←	Malate dehydrogenase, NAD(P)H-binding
				DL693									<i>secY</i> ←	Preprotein translocase membrane subunit
							DL712	DL714	DL698	DL701	DL703	DL707	<i>rplE</i> ←	50S ribosomal subunit protein L5
		DL687	DL692	DL693	DL695								<i>nirB</i> →	Nitrite reductase, large subunit, NAD(P)H-binding
													<i>yhgF</i> →	Predicted transcriptional accessory protein
													<i>prfC</i> ←	Oligopeptidase A
		DL687	DL692	DL693	DL695								<i>hdaD</i> → / → <i>gadE</i>	acid-resistance membrane protein/ DNA-binding transcriptional activator
													<i>gadW</i> ← / → <i>gadX</i>	DNA-binding transcriptional activator/ DNA-binding transcriptional dual regulator
							DL712	DL714	DL698	DL701	DL703	DL707		
		DL687											<i>yhkK</i> ←	Predicted diguanylate cyclase
													<i>yjcm</i> ←	predicted transporter
		DL687	DL692	DL693	DL695								<i>ade</i> →	Cryptic adenine deaminase
				DL693										
													<i>bgfF</i> ←	Fused beta-glucoside specific PTS enzyme
													<i>ompL</i> ←	Predicted outer membrane porin L
		DL687	DL692	DL693	DL695									
							DL712	DL714	DL698	DL701	DL703	DL707	<i>prfA</i> ←	Primosome factor

Table 4.1 continued

MDS pop 1				MDS pop 2			MDS pop 3		MDS pop 4						annotation	gene	description
				DL687	DL692	DL693	DL695								<i>H70F</i> (C <i>A</i> T→C <i>C</i> T)	<i>fabR</i> →	DNA-binding transcriptional repressor
								DL709	DL717						coding (424-432/705 nt)	<i>fabR</i> →	DNA-binding transcriptional repressor
										DL712	DL714	DL698	DL701	DL703	DL707	<i>fabR</i> →	DNA-binding transcriptional repressor
															coding (637/705 nt)	<i>soxS</i> ← / → <i>soxR</i>	DNA-binding transcriptional dual regulator/DNA-binding transcriptional regulator, Fe-S center for redox
															Intergenic (~1185)	<i>mfa</i> →	Nitrite reductase, formate-dependent, cytochrome
				DL687	DL692	DL693	DL695								L325V (L <i>T</i> G→ <i>G</i> TG)		Prolineglycine betaine transporter
						DL693									L51* (T <i>I</i> A→T <i>G</i> A)	<i>proP</i> →	Aspartate ammonia-lyase
								DL709	DL717						E213* (G <i>A</i> A→ <i>I</i> AA)	<i>aspa</i> ←	
												DL698		DL703	DL707	<i>ghvV</i> → / → <i>ghX</i>	RNA-Gly/RNA-Gly
															Intergenic (+11/-26)	<i>ghvV</i> → / → <i>ghX</i>	RNA-Gly/RNA-Gly
												DL698		DL703	DL707	<i>ghvV</i> → / → <i>ghX</i>	RNA-Gly/RNA-Gly
															Intergenic (+18/-19)	<i>ghvV</i> → / → <i>ghX</i>	RNA-Gly/RNA-Gly
												DL698		DL703	DL707	<i>ghvV</i> → / → <i>ghX</i>	RNA-Gly/RNA-Gly
															Intergenic (+21/-16)	<i>rpII</i> →	50S ribosomal subunit protein L9
				DL687											F129V (L <i>T</i> C→ <i>G</i> TC)		Valyl-tRNA synthetase
							DL695								S184S (T <i>C</i> T→T <i>C</i> G)	<i>valS</i> →	

Table 4.1 continued

*Table 4.1 continued* Strains are grouped by population. Annotations denote synonymous (green), nonsynonymous (blue), and nonsense mutations (red). Mutations in intergenic regions are designated by the location of the mutation in relation to coding genes flanking either side. In the case of large deletions, the deleted genes are specified. For changes greater than a single base-pair, the site of the mutation is indicated, followed by the total length of the open reading frame. For example, the “coding (321/1218 nt)” mutation occurs after the 321st base of the 1218 nucleotide reading frame of this gene. Gene orientations are indicated by arrows. Gene descriptions were obtained from EcoCyc.

## Chapter 5: Conclusion and Future Directions

Engineered biological systems are becoming more commonplace solutions to societal challenges. One of my favorite examples, is the microbial production of a protease, chymosin. Hard cheeses are traditionally made using rennet, an ingredient that is harvested by slaughtering young calves and extracting the inner lining of their fourth stomach. The rennet of young calves contains high quantities of chymosin, the main coagulating agent that is used to separate milk into solid casein curds. Chymosin is only present in young calves, it aids them in digesting milk from their mothers. In the 1970s, there was an increasing demand for cheese, and given the dearth of rennet from young calves, a logical alternative was to produce calf chymosin in a microbial host. Fermentation via a microbial host produces large quantities of chymosin. After fermentation, chymosin is extracted and the host is killed to avoid any cross-contamination with the downstream cheese-making process. Today, almost all hard cheeses are made using recombinantly synthesized chymosin with products such as CHY-MAX and MAXIREN that rely on *Aspergillus niger* and *Kluyveromyces lactis* as hosts, respectively. Compared to rennet-derived chymosin, recombinant chymosin is cheaper, uncontaminated (animal rennet contains other enzymes), and leads to more consistent flavors during cheese-making<sup>159</sup>.

Production of a single protein in a microbial host is generally straightforward. However, not all engineering of biological systems is quite so simple. Host organisms are known to mutate in order to adapt to the metabolic load imposed by the engineered function, making extensive engineering efforts challenging. Therefore, in order to make biological engineering easier and more reliable, we need to build genetically stable host organisms and understand the mechanisms involved in evolutionary stability. The complexity of genome engineering efforts is increasing, and evolutionary stability will become more important as more complex functions are encoded in microbial hosts. The

host cell burden will increase causing an increase in mutation rates as complex functions require more DNA. An important aspect of this dissertation is that it describes both the engineering of evolutionary stability and the cellular mechanisms responsible.

Genetically stable strains, also known as antimutators, are difficult to study because their isolation proves challenging. This dissertation demonstrates a new method, Periodic Reselection of Evolutionarily Stable Variants (PResERV), in which antimutators are selected via directed evolution. By performing PResERV in *E. coli*, I have been able to isolate and characterize a series of antimutator strains. Moreover, my work was not an exhaustive search for *E. coli* antimutators, although PResERV could easily be applied to perform such an investigation. PResERV is a universal method that can likely be applied in most microbes. Nearly all antimutators in the literature are reported in *E. coli* and *Saccharomyces cerevisiae*, and therefore there exists an untapped potential for discovering antimutator mechanisms in non-model organisms. Are the antimutator mechanisms similar? Different? Why? This knowledge will allow us to better understand genetic stability and what sorts of cellular processes perturb the integrity of the genome.

This dissertation describes a number of antimutator alleles in the following genes: *polA*, *polB*, *sucD*, *sdhA*, and *rne*. Antimutator alleles in *polA* and *polB* have been previously reported. These alleles support two of the most well-documented antimutator mechanisms. First, DNA polymerase variants with increased proofreading or fidelity (*polA*) and second, deletion of error-correcting DNA polymerases (*polB*) lead to lower mutation rates. Given that *polA* and *polB* have been previously implicated in an antimutator phenotype, they act as a positive control for PResERV and essentially support that PResERV can identify bona fide antimutators. Additionally, this dissertation also expands our knowledge on new antimutator mechanisms by linking mutations in *sucD*, *sdhA*, and *rne* to reduced mutation rates. Succinyl-CoA synthetase, *sucD*, and succinate dehydrogenase, *sdhA*, are components



of central carbon metabolism (TCA cycle and electron transport chain). The antimutator alleles are deletions of these genes and their respective enzyme complexes, which lead to reduced oxidative stress, less DNA damage, and lower mutation rates. However, it still remains unclear why oxidative stress mitigation, a general cellular mechanism, leads to reduced mutation rates on a plasmid versus the chromosome. Future work will be needed to address this question. RNase E, *rne*, is involved in a myriad of processes dealing with RNA metabolism. More follow up studies, such as RNA-Seq are needed in order to tease apart a specific role. This dissertation reports on many other putative antimutator alleles for which the molecular mechanisms of action remain to be elucidated.

Antimutator alleles can be further characterized by their mutational spectrum. There exist novel, sequence-based methods, such as maximum-depth sequencing (MDS), to capture the nature of the mutations found in a given reporter gene<sup>160</sup>. Comparing the antimutator mutational spectrums with wild type will allow us to understand which mutations are corrected by the antimutator allele. For example, in the case of *sdhA* and *sucD*, we would expect that these cells have fewer GC to TA transversions caused by the main species generated during oxidative stress, 8-oxo-G. Additionally, MDS will allow us to calculate mutation rates from the sequence data. These data will be further support for the reductions in mutations rates observed via the Luria- Delbrück fluctuation assay.

Using antimutators as host strains for reliable engineering of biology is a new strategy that is presented extensively in this dissertation. Antimutator alleles (*sucD* and *sdhA*) can be ported to reduce mutation rates in another *E. coli* strain, primarily used for plasmid-based protein production (BL21). In addition, antimutator alleles (*sucD*, *sdhA*, and *rne*) can be combined to reduce mutation rates even further and build a Total Antimutator Combination Organism (TACO). These efforts are the nascent stages of creating antimutator host variants. Next steps consist of porting antimutator alleles into other

organisms to diagnose conservation of antimutator mechanisms across species and building different versions of TACO to understand which (and why) antimutator alleles function in combination. Additionally, it is important to address whether these antimutator alleles will reduced mutation rates in other conditions. For example, the PResERV experiments described in this dissertation were evolved in LB media using shake flasks. Will these mutations work in an industrial process? And, what trade-offs exist with lowering mutation rates? It is likely, at least in the case for the TCA cycle mutants, that the antimutator will have slower growth on certain carbon sources given that the antimutator mutations elicit a diversion of flux through the TCA cycle. Other antimutator alleles, such as those involved in DNA replication, may have different trade-offs. Understanding these limitations will provide a better picture of whether the antimutator phenotypes will transfer to different scenarios. An almost immediate follow-up question that is rarely addressed in this dissertation is – what sorts of synthetic devices will benefit from an antimutator host? What is the relationship between different fold reductions in mutation rate and the cost of a heterologous construct? There is some mathematical modeling that demonstrates that increased production load and escape rate results in a generally rapid appearance of mutated, non-producer cells<sup>23</sup>. It will be important to assay the evolutionary failure of constructs with differing characteristics such as a biosynthesis pathway, a toxic protein, or a circuit. Evolutionary failure is commonly experienced in industrial fermentation processes. For example, engineered strains in a large-scale fermentation will cease to make product after a certain number of generations, due to mutations alleviating the imposed metabolic load of the engineered function. Ultimately, I hope this work will have some translation into an industrial strain to benefit issues faced in industry.

## **Appendix A: Innovation in an *E. coli* evolution experiment is contingent on maintaining adaptive potential until competition subsides**

**This chapter is reproduced (with minor modifications) from its initial publication:**

**Leon D**, D'Alton S, Quandt EM, Barrick JE. (2018) Innovation in an *E. coli* evolution experiment is contingent on maintaining adaptive potential until competition subsides. *PLoS Genetics* **14**, e1007348.

### **Research Contributions**

Leon D, Quandt EM, and Barrick JE conceptualized the work. Leon D and D'Alton S performed experiments. Leon D, D'Alton S, and Barrick JE analyzed data. Leon D and Barrick JE wrote the majority of the manuscript.

### **Acknowledgements**

We thank Zachary Blount, Noah Ribeck, and Michael Wiser for helpful discussions and Craig Barnhart for assistance with genome sequencing. We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources.

## ABSTRACT

Key innovations are disruptive evolutionary events that enable a species to escape constraints and rapidly diversify. After 15 years of the Lenski long-term evolution experiment with *Escherichia coli*, cells in one of the twelve populations evolved the ability to utilize citrate, an abundant but previously untapped carbon source in the environment. Descendants of these cells became dominant in the population and subsequently diversified as a consequence of invading this vacant niche. Mutations responsible for the appearance of rudimentary citrate utilization and for refining this ability have been characterized. However, the complete nature of the genetic and/or ecological events that set the stage for this key innovation is unknown. In particular, it is unclear why it took so long for citrate utilization to evolve and why it still has evolved in only one of the twelve *E. coli* populations after 30 years of the Lenski experiment. In this study, we recapitulated the initial mutation needed to evolve citrate utilization in strains isolated from throughout the first 31,500 generations of the history of this population. We found that there was already a slight fitness benefit for this mutation in the original ancestor of the evolution experiment and in other early isolates. However, evolution of citrate utilization was blocked at this point due to competition with other mutations that improved fitness in the original niche. Subsequently, an anti-potential genetic background evolved in which it was deleterious to evolve rudimentary citrate utilization. Only later, after further mutations accumulated that restored the benefit of this first-step mutation and the overall rate of adaptation in the population slowed, was citrate utilization likely to evolve. Thus, intense competition and the types of mutations that it favors can lead to short-sighted evolutionary trajectories that hide a stepping stone needed to access a key innovation from many future generations.

## **AUTHOR SUMMARY**

Key innovations are rare, game-changing moments in evolution when a species or population achieves new success by escaping its normal constraints. We examined a case in which bacteria that had been maintained in the laboratory for fifteen years evolved to exploit a previously untapped nutrient in their environment. Why didn't this highly beneficial innovation evolve earlier? We found that two distinct mechanisms suppressed this innovation at different times in the history of the population. Early on, competition drove any new cells that started on the path to evolving the innovation extinct. Later, genetic changes accumulated in the population that shut down the potential to benefit from the new nutrient. After competition abated somewhat and further genetic changes restored a beneficial path to the innovation, it evolved. This example illustrates how stiff competition can force evolving populations to adopt short-sighted, incremental solutions that block or significantly delay achieving innovative breakthroughs.

## **INTRODUCTION**

*Escherichia coli* can more effectively uptake iron when a small amount of citrate, about 10  $\mu\text{M}$ , is added to chemically defined media<sup>161</sup>. In 1950, when DM (Davis-Mingioli) medium was initially designed, the recipe used 1700  $\mu\text{M}$  citrate. DM medium was formulated to isolate auxotrophic mutants by the penicillin method. It was shown that the addition of citrate improved the lethality of penicillin, thereby reducing the number of false-positives when selecting for auxotrophs. The penicillin method was widely used, and DM was adopted as a chemically defined medium for other types of experiments as a consequence<sup>162</sup>. The elevated concentration of citrate was typically unnecessary in these new circumstances, but it remained unaltered as new labs and generations of scientists inherited the same DM recipe.

The Lenski long-term evolution experiment (LTEE) consists of twelve *E. coli* B populations that have been propagated daily in glucose-limited DM medium for over 60,000 generations<sup>163,164</sup>. A relatively low concentration of glucose (139  $\mu$ M) was used in the LTEE to restrict the cell density and thereby reduce the chances that stable ecology reinforced by cross-feeding interactions would evolve, which has largely been the case except for in one population<sup>165,166</sup>. The low-glucose formulation of DM used in the LTEE means that the standard amount of citrate present (1700  $\mu$ M) represents a substantial nutrient pool that could be exploited, but the ancestral strain of *E. coli* is unable to utilize citrate under the conditions of the LTEE. Citrate is an untapped niche.

After 31,500 generations, a mutant in one of the LTEE populations, designated Ara-3, evolved the ability to utilize citrate as a carbon source. Descendants of this mutant that were able to fully exploit this additional carbon source evolved by 33,000 generations and dominated thereafter<sup>167</sup>. Previous work has demonstrated that the evolution of citrate utilization in the LTEE proceeded through three stages typical of any key innovation: potentiation, actualization, and refinement<sup>167-170</sup>. The principal mutations involved in the latter two steps have been characterized. Actualization refers to the first manifestation of a rudimentary Cit<sup>+</sup> phenotype. The actualizing mutation is a tandem duplication of the *rnk-citG* region of the *E. coli* chromosome that includes the citrate:succinate antiporter gene, *citT*. This duplication results in an arrangement in which one of the two copies of *citT* is now downstream of the aerobically-active *rnk* promoter ( $P_{rnk}$ ). Thus, the actualization step results in CitT production under the LTEE conditions<sup>168</sup>.

However, early Cit<sup>+</sup> cells with the *citT* duplication are able to uptake and metabolize only a small fraction of the citrate present during one 24-hour growth cycle of the LTEE<sup>167,171</sup>. Subsequently, refinement mutations improved the rudimentary Cit<sup>+</sup> trait in their descendants such that they became capable of utilizing all of the citrate in DM during

each growth cycle (Cit<sup>++</sup> phenotype). One critical mutation for refinement activated expression of *dctA*, a C<sub>4</sub>-dicarboxylate:H<sup>+</sup> symporter gene. DctA allows active transport of C<sub>4</sub>-dicarboxylates, including succinate, into the cell. Because CitT is an antiporter that couples export of these compounds to citrate import, expression of DctA creates a sustainable cycle for importing citrate that is powered by the proton gradient<sup>169</sup>. The *dctA* mutation refines the rudimentary Cit<sup>+</sup> trait into the Cit<sup>++</sup> phenotype that was responsible for the population expansion observed at ~33,000 generations in the LTEE<sup>167</sup>.

While the actualization and refinement stages of Cit<sup>+</sup> evolution in the LTEE are understood, the mechanistic basis of potentiation has remained elusive. The critical diagnostic characteristic of a potentiated strain is that it has an increased chance of giving rise to a Cit<sup>+</sup> descendant after further evolution<sup>168</sup>. Blount *et al.* identified potentiated strains by performing ‘replay’ experiments<sup>167</sup>. In these experiments, pre-Cit<sup>+</sup> clones isolated from the LTEE population at various time points were tested to determine whether they were capable of evolving citrate utilization. Cit<sup>+</sup> cells rarely arose in these replay experiments. When they did, the Cit<sup>+</sup> trait re-evolved more often in clones selected from later time points that were closer to when the *citT* duplication first arose in the LTEE population.

The phylogenetic distribution of the LTEE strains giving rise to Cit<sup>+</sup> variants in the replay experiments suggests the existence of at least two critical junctures at which the potential for evolving Cit<sup>+</sup> increased<sup>168</sup>. By 20,000 generations, the LTEE population had diversified into three long-lived clades that co-existed at least until full citrate utilization (Cit<sup>++</sup>) evolved at ~33,000 generations. *E. coli* isolates from all of these groups evolved Cit<sup>+</sup> in the replay experiments, whereas no strains from earlier than 20,000 generations did, suggesting that all three clades share some determinant of potentiation. A significantly higher proportion of clones from the clade that gave rise to citrate utilization in the LTEE were able to evolve Cit<sup>+</sup> in the replays, suggesting that they share a second determinant for

increased potentiation not present in the other clades. Due to the extreme rarity of Cit<sup>+</sup> arising in these replay experiments even after months of evolution<sup>167</sup>, it is not realistic to use this approach to further narrow down the genetic basis of potentiation.

In this study, we tested the viability of the critical actualizing mutation for Cit<sup>+</sup> evolution in a series of pre-Cit<sup>+</sup> isolates from the LTEE by measuring the effect of activating *citT* expression on competitive fitness. We found that activating *citT* expression slightly increased the fitness of the ancestral strain and some later pre-Cit<sup>+</sup> clones. Unexpectedly, activating *citT* expression was highly deleterious in certain strains from intermediate time points, and we did not find any strains that benefitted significantly more from this mutation than the ancestral strain did. We conclude that potentiation for the evolution of citrate utilization in the LTEE is due to the interplay of genetic factors in specific strains and the population at large. First, adaptation had to occur via a genetic trajectory that maintained the potential for evolving Cit<sup>+</sup> by a beneficial mutational step in order for the innovation to remain accessible. Second, the rate of adaptation in the overall population needed to slow to a pace at which early variants with the weakly beneficial Cit<sup>+</sup> trait could avoid being driven extinct by competitors before refining mutations arose.

## RESULTS

### **Cit<sup>+</sup> was only slightly beneficial when it evolved in the LTEE**

By definition, the first *E. coli* cell that evolved the *citT*-activating mutation that was ultimately successful in the LTEE was fully potentiated when this mutation arose. The earliest Cit<sup>+</sup> descendant of this cell that has been identified is strain ZDB564 from 31,500 generations. At this time Cit<sup>+</sup> cells were still extremely rare in the population<sup>170</sup>, which means that it is likely that the suite of mutations in ZDB564 is identical to those in the first



Cit<sup>+</sup> cell, or nearly so. Previously, strain ZDB706, a Cit<sup>-</sup> revertant of ZDB564, was isolated by passing ZDB564 on DM medium lacking citrate to allow for the spontaneous collapse of the *rnk-citG* duplication to the ancestral single-copy state that lacks a copy of the *rnk* promoter upstream of *citT* (**Figure A1A**)<sup>170</sup>.

We co-cultured ZDB564 and ZDB706 in DM medium to estimate the effect that the *citT* duplication had on competitive fitness when it originally arose. These experiments involved reverting an arabinose-utilization allele in one of the two strains to be competed from the inactivated state present in all strains from this LTEE population (Ara<sup>-</sup>) to the active state (Ara<sup>+</sup>) so that cells of each type can be distinguished by the colors of the colonies that they form on indicator plates (**Methods**)<sup>164</sup>. These Ara<sup>+</sup> strain variants were assayed to establish that the genetic marker was neutral with respect to fitness and that no secondary mutations affecting fitness had accumulated during strain construction prior to further competition experiments. When competing ZDB564 and ZDB706, we found a slight fitness advantage of 2.2% for the presence of the *citT*-activating duplication in ZDB564 (**Figure A1B**). This result that was consistent between the competitions utilizing Ara<sup>+</sup> ZDB706 or Ara<sup>+</sup> ZDB564 marked variants (two-tailed *t*-test, *P*=0.33, *n*=12 and 18, respectively).

### Development of P<sub>*rnk-citT*</sub> knock-in assay for potentiation

We next wanted to add the *citT*-activating mutation to pre-Cit<sup>+</sup> strains in order to test our hypothesis that there was a transition in the lineage leading to Cit<sup>+</sup> such that this mutation became more beneficial once a potentiated genetic background evolved. The effect of adding a plasmid containing the evolved P<sub>*rnk-citT*</sub> unit has been tested in previous studies<sup>168,169</sup>, but this approach is problematic because these plasmids are multicopy,

whereas only a single activated copy of the *citT* gene was present in the initial Cit<sup>+</sup> strains. However, engineering the authentic *rnk-citG* duplication into the chromosome of a strain is difficult because this configuration is genetically unstable. It readily collapses via homologous recombination if there is not selection to maintain citrate utilization, as was utilized in reverting ZDB564 to the Cit<sup>-</sup> variant ZDB706.

To address these shortcomings, we developed a P<sub>*rnk-citT*</sub> knock-in assay, in which a mimic of the evolved configuration is integrated into the chromosome of a pre-Cit<sup>+</sup> LTEE clone (**Figure A1C**). Briefly, we created an activated *citT* module linked to an antibiotic selection marker in which the *rnk* promoter is upstream of the truncated *rnk-citG* fusion ORF formed by the duplication followed by the complete *citT* reading frame. To control for any fitness cost imposed by the selection marker, we also made a null module containing only the antibiotic resistance gene. Both of these cassettes are targeted to integrate into the *E. coli* chromosome such that they replace the *lac* operon, which is unrelated to citrate or glucose metabolism.

We validated this approach by adding the P<sub>*rnk-citT*</sub> module to the fully potentiated Cit<sup>-</sup> revertant, ZDB706, and adding the null module to its neutral Ara<sup>+</sup> variant. Addition of the P<sub>*rnk-citT*</sub> cassette to the fully potentiated Cit<sup>-</sup> strain ZDB706 resulted in increased *citT* mRNA levels equivalent to those seen in ZDB564, the original Cit<sup>+</sup> isolate with the actual *rnk-citG* duplication that evolved in the LTEE (two-tailed *t*-test,  $P = 0.65$ ,  $n = 3$ ) (**Figure A1D**). The resulting Cit<sup>+</sup> variant of ZDB706 had a fitness advantage of 2.4% over the corresponding Cit<sup>-</sup> variant with the null knock-in cassette (**Figure A1B**), which was not statistically different from the fitness advantage found for the authentic *citT*-activating mutation in the pooled ZDB564 versus ZDB706 competitions (two-tailed *t*-test,  $P = 0.71$ ,  $n = 6$  and 30, respectively). Therefore, applying the P<sub>*rnk-citT*</sub> knock-in assay to additional

strains allows us to ask: if the *citT*-activating mutation had evolved in a genetic background that existed earlier in the LTEE, would it have been as beneficial?

### **Cit<sup>+</sup> would have been modestly beneficial if it evolved in the LTEE ancestor**

As a first step in further elucidating the fitness consequences of evolving rudimentary Cit<sup>+</sup> on other strains from the LTEE, we performed the  $P_{rnk}$ -*citT* knock-in assay on the ancestral LTEE strain, REL606. We found a slight fitness benefit of 1.0% for the Cit<sup>+</sup> mutation (**Figure A1B**). This effect size is near the limit for the smallest differences that can be distinguished in these types of competitive fitness assays, resulting in relatively weak support for the hypothesis that there was any fitness advantage at all for the REL606 variant with the  $P_{rnk}$ -*citT* module relative to the one with the null module (one-tailed *t*-test,  $P = 0.033$ ,  $n = 12$ ). There was evidence, though also not very strong, that the benefit of the  $P_{rnk}$ -*citT* module in the fully potentiated strain ZDB706 was greater than it was in REL606 (one-tailed *t*-test,  $P = 0.018$ ,  $n = 12$  and 6, respectively). Expression of *citT* was not quite as high in the REL606 strain with the  $P_{rnk}$ -*citT* module as it was in ZDB706 with the same module (two-tailed *t*-test,  $P = 0.00016$ ,  $n = 3$ ) (**Figure A1D**), suggesting that mutations during the LTEE on the lineage leading to Cit<sup>+</sup> may have altered the strength of the *rnk* promoter. Overall, the REL606 measurements indicated, surprisingly, that there was likely a modest benefit for a mutation activating expression of *citT* at the very beginning of the LTEE, and that this benefit may have only slightly improved after further mutations that occurred during the potentiation stage in the evolution of this metabolic innovation.

### **No evidence for ecological potentiation**

Why did the appearance of citrate utilization take so long and why has it not evolved in other LTEE populations? One hypothesis for its rarity is that the evolution of a particular ecology in the population was important for enabling the evolution of Cit<sup>+</sup>. This type of situation is known to occur, for example, when nutrient cross-feeding between genetically diverged subpopulations yields negative frequency dependence, such that the competitive advantage for a newly evolved strain or a certain subpopulation is greater when it is rare within the population than when it is common<sup>73</sup>. The pre-Cit<sup>+</sup> clade was rare during the time period when the *rnk-citG* duplication evolved. It constituted <1-5% of the population from 30,000 to 32,500 generations<sup>170</sup>.

To test whether this kind of ‘ecological potentiation’ was important for the evolution of Cit<sup>+</sup> in the LTEE, we repeated the *P<sub>rnk-citT</sub>* knock-in assay competition for strain ZDB706 in the context of the full diversity that existed in the population at 31,000 generations (**Figure A1B**). The Cit<sup>+</sup> and Cit<sup>-</sup> variants were mixed together equally and added such that they comprised ~1% of the cells in a mixture with the evolved population sample. In this context, the Cit<sup>+</sup> strain had a 0.9% fitness advantage over the Cit<sup>-</sup> strain, which was less than and only marginally different from the result when the two strains were competed versus one another normally (two-tailed *t*-test, *P* = 0.053). Thus, we find no support for the ecological potentiation hypothesis. If anything, the more diverse mixed population context may slightly reduce the benefit of Cit<sup>+</sup> evolution.

### **Anti-potentiated strains evolved at intermediate time points**

We next performed the *P<sub>rnk-citT</sub>* knock-in assay on 23 additional clones isolated from the LTEE population (**Figure A2A**). Our goal was to determine whether activating *citT*

expression was similarly beneficial in other evolved genetic backgrounds. We measured *citT* mRNA levels in five of the constructed strains with the  $P_{rnk}$ -*citT* cassette and found them to be similar in all of these strains, indicating that the strength of the *rnk* promoter was largely unchanged by the specific suites of evolved mutations present in each of these strains. For four of the evolved strains we found strong evidence that the *citT* cassette significantly increased fitness versus the control with the null cassette, as it had in the fully potentiated strain ZDB706 (one-tailed bootstrap test incorporating Ara<sup>+</sup>/Ara<sup>-</sup> marker and Cit<sup>+</sup>/Cit<sup>-</sup> competitions described in **Methods**,  $P < 0.05$ ). In nine strains, *citT* activation had no significant effect on fitness (two-tailed bootstrap test,  $P < 0.05$ ), though our measurements did not achieve sufficient precision to rule out that there was a fitness benefit of 1% or greater in seven of these cases (one-tailed bootstrap test,  $P < 0.05$ ).

Unexpectedly, the Cit<sup>-</sup> variant outcompeted the Cit<sup>+</sup> variant for the 11 remaining strains of the 23 we tested (one-tailed bootstrap test,  $P < 0.05$ ). The actualizing step needed for subsequently evolving full citrate utilization (Cit<sup>++</sup>) would have been effectively blocked if it occurred in these strain backgrounds; they are ‘anti-potentiated’. For five of these strains, activating *citT* expression was extremely detrimental, decreasing competitive fitness by >20% (one-tailed bootstrap test,  $P < 0.05$ ). For ZDB483 and ZDB14, two of the severely anti-potentiated strains, we investigated the nature of this defect by comparing growth curves of the Cit<sup>+</sup> and Cit<sup>-</sup> variants. There was very little difference in the growth curve for the LTEE ancestor REL606 whether the activated *citT* cassette or null control cassette was added to its genome, which is in keeping with its almost imperceptible effect on the competitive fitness of this strain. In contrast, we found that activating *citT* expression drastically increased the lag phase of growth in the severely anti-potentiated strains (**Figure A2B**). This additional lag time can explain the sizable competitive disadvantage versus the

Cit<sup>-</sup> strain, even though the Cit<sup>+</sup> variants are able to reach a higher final cell density if cultured alone.

### Mapping potentiation onto phylogeny

Identifying specific mutations that contributed to potentiation and anti-potentiation requires interpreting the fitness data from the *P<sub>rmk</sub>-citT* knock-in assays in a phylogenetic context. To improve the resolution of a previously published whole-genome phylogenetic tree of 29 clonal isolates from this LTEE population<sup>168</sup>, we sequenced the genomes of 20 new clones (**A1 Table**) and also incorporated 12 other clones sequenced in another recent study of the rate of genome evolution through 50,000 generations in all LTEE populations<sup>83</sup>. The 20 newly sequenced isolates were selected to improve our ability to temporally order mutations that occurred near when citrate utilization evolved: they were minimally diverged from the line of descent to the Cit<sup>+</sup> progenitor and were mostly sampled at later time points.

The updated phylogenetic tree (**Figure A3**) includes all 25 clones we tested with the *P<sub>rmk</sub>-citT* knock-in assay. We used these strains to identify branches in the tree within which the adaptive potential of activating *citT* expression changed due to one or more mutations. Specifically, we clustered phylogenetically-adjacent strains into groups within which all pairwise comparisons of the fitness effect of the *P<sub>rmk</sub>-citT* module were not significantly different (Bonferroni-corrected two-tailed bootstrap tests,  $P > 0.05$ ). Overall, this analysis suggests that there were at least three major step-like changes in the potential for evolving the rudimentary Cit<sup>+</sup> trait along the pre-Cit<sup>-</sup> lineage that eventually evolved citrate utilization (**Figure A4**).

Proceeding backward in the tree from the earliest known Cit<sup>+</sup> isolate (ZDB564), two earlier clones (ZDB19 and ZDB13) from as early as 29,000 generations are as fully potentiated as the key Cit<sup>-</sup> revertant (ZDB706). The overall fitness effect of evolving Cit<sup>+</sup> in this group was +2.4% [+1.4%, +3.4%] (95% confidence interval). The next-earliest group comprises three clones isolated at time points from 25,000 to 27,000 generations (ZDB478, ZDB486, and ZDB309). Activation of *citT* had little to no impact on this set of strains, with an estimated group-wise effect on fitness of +0.4% [-1.3%, 2.0%]. One intermediate strain, ZDB310 from 27,000 generations, was not significantly different from either of these two groups immediately before and afterward, although the two groups were significantly different from one another.

It was deleterious to evolve Cit<sup>+</sup> in an earlier, intermediate set of isolates composed of ZDB425, ZDB458, and ZDB464 with an estimated fitness effect of -5.4% [-7.2%, -3.6%]. These clones appear to be genetically typical of the pre-Cit<sup>+</sup> lineage. ZDB425 at 10,000 generations and ZDB458 at 20,000 generations have only one and two ‘private’ mutations not shared with the main pre-Cit<sup>+</sup> lineage, respectively, though we cannot rule out that other changes in the impact of *citT* activation may have occurred on the main line of descent within this interval. Before these anti-potentiated clones, there is an initial cluster that groups ZDB409 and ZDB429 with the REL606 ancestor. In these three isolates, evolution of Cit<sup>+</sup> would have been slightly beneficial with a fitness impact of +1.7% [+0.0%, +3.3%].

Other strains are not classified into these major groups. It is less likely that they are representative of how potentiation evolved in the lineage leading to Cit<sup>+</sup>. For example, the four most highly anti-potentiated clones (ZDB467, ZDB483, ZDB14, ZDB18) appear to have evolved this property independently and due to ‘private’ mutations not shared with the main pre-Cit<sup>+</sup> lineage (**Figure A3**), at least this is the most parsimonious explanation.

Similarly, the fitness effects measured in the  $P_{rnk-citT}$  knock-in assay for two three-member subclades (ZDB334, ZDB339, ZDB317; and ZDB23, ZDB27, ZDB25) indicate that each likely shared one or more mutations that altered  $Cit^+$  potentiation only within that subclade, though the effects are much smaller in these cases. Finally, we excluded ZDB446 from this analysis because it was so deeply branched: removed by >5,000 generations from the pre- $Cit^+$  lineage. It would have been clustered with the earliest group containing REL606 according to our criteria.

### **$Cit^+$ evolution in the context of competition with other beneficial mutations**

During the time period when the pre- $Cit^+$  lineage was anti-potentiated, from approximately 10,000 to 20,000 generations, invasion of a new  $Cit^+$  subpopulation would have been nearly impossible. Lineages that lost fitness by evolving the rudimentary version of this new trait would be rapidly purged by selection before refining mutations (e.g., activating *dctA*) could accumulate to give the decisive benefit of full citrate utilization (the  $Cit^{++}$  phenotype). What about the earlier and later time periods when the evolution of  $Cit^+$  was neutral or slightly beneficial? During these epochs, a newly evolved  $Cit^+$  lineage would still have had to compete with not only its own ancestor, but also against other lineages that were evolving at the same time, many of which would have other beneficial mutations. That is, an incipient  $Cit^+$  lineage had to survive in competition with alternative adaptive pathways, such as those improving fitness on glucose.

In order to understand when the fitness effects we measured for evolving  $Cit^+$  by *citT* activation would have made this metabolic innovation a viable evolutionary pathway in the context of competition within the LTEE population, we compared the group-wise fitness effects determined from the  $P_{rnk-citT}$  knock-in assays to two models of the fitness



effects of beneficial mutations that were successful at different generations in this LTEE population (**Figure A4**). The Wiser *et al.* approach fits the fitness trajectory of this LTEE population to a model that incorporates a uniform type of diminishing returns epistasis between beneficial mutations and assumes consecutive sweeps<sup>172</sup>. The Tenaillon *et al.* model fits the number of beneficial mutations accumulating over time from genome sequencing data<sup>83</sup>. We combined this information with the Wiser *et al.* fitness trajectory to infer the representative fitness change for each subsequent beneficial mutation. The larger fitness effects in the Wiser *et al.* model reflect that it estimates the advantages of sweeping cohorts that may include more than one beneficial mutation. Overall, both models give very similar results that reflect the well-known deceleration in fitness gains during the Lenski long-term experiment<sup>163,164,173</sup>.

The models demonstrate that even if Cit<sup>+</sup> evolution was marginally beneficial in the REL606 ancestor and other early isolates, it was initially much less beneficial than was needed to be successful at this point. Even by 5,000 generations, *citT* activation appears to have been average, at best, in terms of its fitness effect among all possible beneficial mutations. It would have been unlikely for the Cit<sup>+</sup> trait to appear and persist at this point because there were so many alternative mutations, such as those that required only single-base substitutions or IS insertions that knocked out gene function, which would have occurred at a higher rate than the specific duplications or IS element insertions needed to activate *citT* expression<sup>167</sup>. After anti-potential appeared and receded in this lineage, competition would have continued to suppress Cit<sup>+</sup> evolution when the *citT* mutation was again neutral. In striking contrast, evolving Cit<sup>+</sup> was clearly superior to a typical successful beneficial mutation in the final group of strains that first evolved by 29,000 generations. It was a viable adaptive pathway at this point. Thus, by comparing P<sub>rnk-citT</sub> knock-in assays to models of the rates of population evolution, we can explain how a variant with a

rudimentary  $\text{Cit}^+$  trait was able to appear and avoid extinction long enough to achieve the decisive *dctA* mutation that led to the dominant  $\text{Cit}^{++}$  trait.

## DISCUSSION

Our work reframes and further elucidates why the emergence of citrate utilization is so rare in the Lenski long-term evolution experiment (LTEE). Rudimentary citrate utilization (the  $\text{Cit}^+$  phenotype) can apparently evolve at any time when a mutation switches on expression of the *CitT* transporter under the aerobic conditions of the experiment. However, the success of a new  $\text{Cit}^+$  variant is far from guaranteed. It is contingent on whether its descendants can survive long enough to incorporate a second mutation, such as one activating expression of the *DctA* transporter, that enables full citrate utilization (the  $\text{Cit}^{++}$  phenotype). The chance that  $\text{Cit}^{++}$  will be realized by this evolutionary pathway is dependent on two major factors. First, the initial mutational step conferring the weak  $\text{Cit}^+$  phenotype must be beneficial to fitness. Whether it is advantageous or not depends on the context of other mutations present in an evolved genome in which *citT* activation occurs. Second, the benefit of the mutation conferring weak  $\text{Cit}^+$  must be great enough that it can survive in competition with other adaptive mutations. Whether it is sufficiently beneficial depends on the population context in which it arises. We found that both genetic and population factors limited  $\text{Cit}^{++}$  evolution at different times in the LTEE (**Figure A4**).

Unexpectedly, evolution of  $\text{Cit}^+$  by activating *citT* expression appears to have already been slightly beneficial to fitness in the ancestral strain used to found this *E. coli* population on the first day of the LTEE and to have remained so in other early evolved isolates. Even though  $\text{Cit}^+$  strains that evolved in the LTEE population at this point would have been capable of displacing their own  $\text{Cit}^-$  ancestors, this first step on the pathway to

the full Cit<sup>++</sup> innovation was suppressed due to competition with mutations on adaptive pathways that improve fitness in the original glucose niche. New cells with highly beneficial mutations related to this primary component of the LTEE environment were essentially guaranteed to arise in the population and outcompete any cells with mutations activating *citT* expression. By 10,000 generations, the lineage in which Cit<sup>+</sup> eventually evolved became ‘anti-potentiased’ after it accumulated additional mutations. Now, the pathway to innovation was blocked because it was deleterious to evolve rudimentary Cit<sup>+</sup> in this genetic background. There was a fitness valley separating the evolved Cit<sup>-</sup> strains from the full Cit<sup>++</sup> phenotype. Finally, further mutations appeared in the focal LTEE lineage by 29,000 generations that altered the fitness impact of activating *citT* expression such that it was again beneficial to evolve the Cit<sup>+</sup> phenotype, and perhaps even more so than it had been in the ancestor. At this point, the rate of adaptation of the population had slowed enough that evolving rudimentary Cit<sup>+</sup> was now among the most beneficial mutational steps remaining. The two-step mutational pathway to Cit<sup>++</sup> was no longer suppressed by genetic or population factors, and the Cit<sup>++</sup> innovation evolved.

Cit<sup>++</sup> mutants of *E. coli* capable of growth on citrate as a sole carbon source under aerobic conditions have been isolated in other studies<sup>167,168,174,175</sup>. In all of these cases, multiple mutations have been required to achieve the Cit<sup>++</sup> phenotype. When they have been identified, the mutations that yield Cit<sup>++</sup> activate expression of the CitT and DctA transporters, as is observed in the LTEE. These studies have isolated Cit<sup>++</sup> mutants in much shorter periods of time (<1-8 weeks) than it took to evolve in the LTEE (~15 years) because they involve starving *E. coli* cells for days to weeks under conditions in which citrate was present as a potential carbon source. In the context of our results and as previously noted by others<sup>176</sup>, this difference in environmental conditions relative to the glucose-limited transfer regime of the LTEE, in which cells are in stationary phase for only ~16-18 hours

each day, dramatically increases the fitness benefit of evolving the rudimentary Cit<sup>+</sup> phenotype. Therefore, these stark conditions are expected to completely mask and overwhelm the dependency on potentiating genetic and population factors found in the LTEE. Activating *citT* expression would be universally beneficial in any genetic background in these types of experiments. Any increase in lag phase or other trade-off with respect to growth rate that might accompany this intermediate step in the pathway to Cit<sup>++</sup> is irrelevant when cells without the mutation simply cannot replicate at all. The citrate-only starvation conditions also eliminate any interference from alternative mutations with benefits related to glucose utilization that suppress Cit<sup>++</sup> evolution in the LTEE.

Why is evolution of Cit<sup>+</sup> beneficial in some evolved genetic backgrounds and deleterious in others under the conditions of the LTEE? Activation of CitT expression under these aerobic conditions via the *rnk-citG* duplication leads to coupled import of citrate (a C<sub>6</sub>-tricarboxyate) and export of C<sub>4</sub>-dicarboxylates (e.g., succinate)<sup>177</sup>. In wild-type *E. coli* strains, CitT is normally expressed only under anaerobic conditions, and the imported citrate can only be assimilated when a fermentable co-substrate, such as glucose, is also present<sup>178</sup>. Under these conditions, citrate is cleaved to acetate and oxaloacetate by citrate lyase. The structural proteins and accessory factors necessary for producing this enzyme complex are encoded in the same operon as *citT*. When glucose is co-utilized with citrate, the resulting oxaloacetate is reduced to succinate by reverse tricarboxylic acid (TCA) cycle reactions. This process consumes reduced cofactors produced by breakdown of the sugar to balance redox metabolism without the need for O<sub>2</sub>. The succinate or other C<sub>4</sub>-dicarboxylates produced can be exchanged for more citrate import via CitT to continue this mixed fermentation mode of growth, or these TCA cycle intermediates can be siphoned off into biosynthetic pathways as necessary for cellular replication.

Under the aerobic conditions of the LTEE, citrate lyase is not expressed and succinate to balance citrate import by CitT must be produced in a different manner, from citrate or glucose using reactions of central metabolism. The availability of O<sub>2</sub> makes it possible to maintain redox balance while synthesizing succinate via the TCA cycle, the glyoxylate bypass, or anaplerotic reactions (e.g., phosphoenolpyruvate carboxylase). *E. coli* growing under aerobic conditions ferments glucose to acetate, and mutations in genes related to the ability to re-uptake and utilize acetate are widespread in the LTEE<sup>83,179,180</sup>. These mutations affect acetate transporters and also pathways for assimilating acetate as acetyl-CoA through citrate synthase, the TCA cycle, and the glyoxylate bypass. Therefore, how these pathways are altered by adaptation to better utilize glucose and acetate is likely an important determinant of the genetic background that affects the ability to evolve citrate utilization. If introduction of the CitT transport reaction misbalances the redox state of the cell or the distribution of carbon compound intermediates between anabolism and catabolism, then it would be deleterious to fitness. Therefore, mutations altering central metabolism are candidates for explaining the changes in the fitness effect of *citT* activation along the LTEE lineage that ultimately evolved citrate utilization (**Figure A4**).

Starting with the ancestor and examining when changes in the potential for evolving Cit<sup>+</sup> were observed in the LTEE, a mutation in *nadR*, a repressor of NAD coenzyme biosynthesis<sup>181</sup>, occurs along the branch in the phylogenetic tree when anti-potential first evolved, before 10,000 generations. Mutations in *nadR* have appeared and swept to fixation in all twelve LTEE populations. These mutations include frameshift mutations and IS element insertions<sup>83</sup>, indicating that they are loss-of-function mutations, and deleting this gene from the genome of the LTEE ancestor has been shown to be beneficial<sup>173</sup>. Reducing or eliminating NadR activity is predicted to increase the NAD/NADH pool in the cell and could enable increased rates of glucose fermentation. Since NADH is a potent

allosteric regulator of enzymes in central metabolism, including citrate synthase (*gltA*) for entry into the TCA cycle, this mutation may also reconfigure other cellular fluxes in ways that make CitT transport deleterious to fitness.

Between 10,000 and 25,000 generations mutations occurred in this LTEE population in three key genes that affected the activities of enzymes in central metabolism: *iclR*, *arcB*, and *gltA1*. These mutations have all been shown to improve growth on acetate<sup>170</sup>. Two of these mutations are in negative regulators; they are expected to derepress enzymes of the glyoxylate bypass (*iclR*)<sup>182</sup> and TCA cycle (*arcB*)<sup>183</sup>, leading to increased metabolic flux through these pathways. Mutations in both of these genes are found in nearly all LTEE populations<sup>83</sup>. The citrate synthase mutation (*gltA1*) reduces allosteric inhibition of this enzyme by NADH<sup>170</sup>, which increases flux of acetyl-CoA into the TCA cycle. Evolved strains with mutations in *gltA* have only persisted in one other LTEE population that maintained the low ancestral mutation rate through 50,000 generations<sup>83</sup>. Both the *arcB* and *gltA1* mutations occurred on a branch in the phylogenetic tree for the citrate LTEE population when the effect of *citT* activation reverted to being neutral with respect to competitive fitness, so they are candidates for reversing anti-potential. The *iclR* mutation does not seem to have had an effect on genetic potential on its own, but it may have interacted with the *arcB* and/or *gltA1* mutations in a way that contributes to this anti-potential effect.

Only one mutation in a gene known to be involved in central metabolism occurred around 27,000 to 29,000 generations, at the point in the phylogenetic tree when adding the *citT* mutation seems to have again become beneficial to fitness. This mutation is upstream of the *ilv* operon for branched chain amino acid biosynthesis in the *yifB/ilvL* intergenic region. This pathway consumes pyruvate and acetyl-CoA, and its products can be used to synthesize the pantothenate moiety of coenzyme A (CoA)<sup>184</sup>. If this mutation affects gene

expression of the *ilv* operon, then it could impact the balance of citric acid cycle intermediates flowing into or out of the TCA cycle to sustain cellular growth directly or indirectly via changing CoA/acetyl-CoA availability.

While the functions of the genes that we have highlighted in central metabolism suggest that they may be especially important for altering the potential for Cit<sup>+</sup> evolution, other mutations also accumulated on the branches in the phylogenetic tree where the effects of *citT* activation on *E. coli* fitness changed (**Figure A3**). In future work, the *P<sub>rnk</sub>-citT* knock-in assay can be used to further dissect this adaptive pathway by testing strains in which various evolved alleles have been removed or added. As an example of this type of approach, we have previously shown that removing the *gltA1* mutation from the earliest Cit<sup>+</sup> isolate (ZDB564) makes the *citT*-activating duplication highly deleterious because it introduces a growth lag like that observed in the strongly anti-potentiated LTEE isolates in this study<sup>170</sup>. Similar studies could be conducted on strains that represent as closely as possible the genotypes present at critical junctures in the phylogenetic tree to determine which mutations altered the chances of achieving this innovation.

Another remaining question is whether the Cit<sup>+</sup> innovation will ever evolve in the other eleven LTEE populations. It has not as of more than 60,000 generations<sup>180</sup>, nearly twice the amount of time that was required for it to evolve in the population analyzed here<sup>167</sup>. The ‘innovation interference’ of other highly beneficial mutations within a population suppressing Cit<sup>+</sup> evolution has undoubtedly faded in all eleven of these populations as the pace of fitness increase has slowed similarly in all of them<sup>172,185</sup>. However, the ubiquity of *nadR* mutations in the LTEE may indicate that other populations similarly descended into a genetically anti-potentiated state. Our results suggest that Cit<sup>++</sup> may still appear in the future if mutations suitably adjust fluxes in central metabolism to make evolving rudimentary Cit<sup>+</sup> by activating *citT* expression a beneficial step on the

pathway to innovation, as long as no critical components have been irrecoverably lost from the genome. Through 50,000 generations, no population has deleted either *citT* or *dctA*, and these genes have not accumulated any mutations in most populations<sup>83</sup>, so the latent genetic potential to evolve Cit<sup>+</sup> seems to have remained intact so far.

The LTEE is an open-ended evolution experiment<sup>186</sup>; it did not begin with the aim of isolating *E. coli* that utilize citrate. There was never strong selection for this novel capability. Because evolving citrate utilization allowed the new Cit<sup>++</sup> clade to colonize an untapped nutrient niche and rapidly diversify, this new metabolic capacity is an example of a key evolutionary innovation<sup>187</sup>. The evolution of Cit<sup>++</sup> initiated a new round of rapid evolutionary optimization that included mutations that reduced the activity of citrate synthase (*gltA2*) and eliminated flux through the glyoxylate shunt (*aceA*), both of which reversed the effects of pre-Cit<sup>+</sup> adaptive mutations<sup>170</sup>. The many new possibilities for improving fitness in this alternative niche also likely contributed to the evolution of hypermutation within the Cit<sup>++</sup> clade by 36,000 generations<sup>168</sup>. Lastly, new ecological interactions arose in this population such that Cit<sup>-</sup> and Cit<sup>++</sup> types co-existed via negative-frequency dependent interactions for at least 10,000 generations after Cit<sup>++</sup> evolved<sup>167,168</sup>. Continuing evolution of interactions between these and other *E. coli* lineages led to the emergence of an ecology that is unique to this flask in the LTEE<sup>171</sup>.

We found that a metabolic innovation in a laboratory population of *E. coli* was contingent on both a history of genetic adaptation and ongoing population dynamics. Evolution of metabolic capabilities has been found to be crucial to the emergence and continued success of bacterial pathogens in several instances<sup>188,189</sup>. For example, *Salmonella* acquired the ability to use tetrathionate as an electron acceptor, giving it a growth advantage relative to other bacteria in the environment that it creates in the gut during infection by inducing inflammation<sup>190</sup>. On a shorter timescale, mutations in the



opportunistic pathogen *Pseudomonas aeruginosa* that accumulate during chronic infections in the cystic fibrosis lung lead to an increased ability to acquire iron from hemoglobin<sup>191</sup>. Even in the simple environment of the LTEE, both genetic and population factors suppress the evolution of an innovation that allows a new niche to be exploited by a new bacterial species. It may be useful in the treatment of disease to understand when these and other factors, including competition for specific nutrients by commensal species in a microbiome, can be used to suppress evolutionary outcomes that are harmful to human health<sup>45</sup>.

## **MATERIALS AND METHODS**

### **Media conditions and strains**

*E. coli* were cultured in Davis-Mingioli (DM) medium and Lysogeny Broth (LB)<sup>170</sup>. As necessary, media were supplemented with 50 µg/mL kanamycin and 80 µg/mL 5-bromo-4-chloro-3-indolyl β-d-galactopyranoside (X-gal). Evolved clones characterized in this study from archived LTEE populations and strain ZDB706 (the spontaneous Cit<sup>-</sup> revertant of ZDB564) were isolated in previous studies<sup>167,168,170</sup>. New strains constructed in this study are listed in **Table A2**.

### **P<sub>rnk-citT</sub> knock-in assay**

The activated P<sub>rnk-citT</sub> module was constructed by amplifying the evolved *rnk-citG* duplication junction from the pCit plasmid along with a linked kanamycin resistance gene (Kan<sup>r</sup>)<sup>169</sup>. The P<sub>rnk-citT</sub> construct in pCit is originally from evolved strain CZB154<sup>169</sup>. As a control, another module was created which only contains the Kan<sup>r</sup> marker. These modules

were integrated into the genomes of several Cit<sup>-</sup> strains (REL607, REL1166A, ZDB429, ZDB467, and ZDB483) via lambda Red recombination<sup>111</sup> such that they replaced the *lac* locus (*lacA* to *lacZ*), spanning positions 333,862-337,485 in the REL606 genome (GenBank:NC\_012967.1)<sup>130</sup>. We transferred the cassettes to other strains using P1 bacteriophage transduction<sup>110</sup>. Successful transductants were scored based on blue/white screening in the presence of X-gal and kanamycin. All Cit<sup>+</sup> strains were made by transduction of the P<sub>*rnk-citT*</sub> module into an Ara<sup>-</sup> LTEE clone. Isogenic Cit<sup>-</sup> strains were constructed by insertion of the control Kan<sup>r</sup> module into an Ara<sup>+</sup> version of the same clone generated as described in the next section. To determine whether any other mutations present in the evolved strains from the LTEE were altered during transduction, we screened for mutations identified by whole-genome sequencing in the recipient strain that were within 100 kb upstream or downstream of the P<sub>*rnk-citT*</sub> insertion site. Strains from three Cit<sup>+</sup>/Cit<sup>-</sup> pairs were found to have gained or lost evolved alleles in this process.

### **Selection for spontaneous Ara<sup>+</sup> mutants**

All Ara<sup>-</sup> strains inherited a point mutation in *araA* present in the REL606 LTEE ancestor that prevents arabinose utilization<sup>173</sup>. To isolate spontaneous Ara<sup>+</sup> mutants, Ara<sup>-</sup> strains were revived overnight at 37°C in DM containing 1 mg/mL glucose (DM1000). For each strain, three separate flasks containing 10 ml of DM1000 were each inoculated with ~500 cells from the first DM1000 culture to reduce the chance that they might share any secondary mutations affecting fitness. After incubating overnight at 37°C, cells were harvested by centrifugation at 4,000 rpm for 15 min and the entire volume was plated on minimal arabinose (MA) plates. Plates were incubated for 36-48 h and colonies were streaked and grown on new MA plates before picking single-colonies as candidate Ara<sup>+</sup>

revertants. The presence of secondary mutations affecting fitness was assessed by competing the original Ara<sup>-</sup> and selected Ara<sup>+</sup> strains, as described below. In most cases, we identified an Ara<sup>+</sup> revertant with a fitness that was not significantly different from its Ara<sup>-</sup> progenitor.

### **Relative fitness measurements**

Relative fitness was measured using co-culture competition assays<sup>164,192</sup>. Two strains to be competed are differentiated based on their ability to ferment arabinose. Ara<sup>-</sup> strains form red colonies on tetrazolium arabinose (TA) media, and Ara<sup>+</sup> strains form pink colonies. Strains were revived overnight in LB then were diluted 10,000-fold into separate cultures for each replicate competition assay in DM containing 25 µg/mL glucose (DM25). These cultures were preconditioned and competed under the same conditions as used in the LTEE<sup>164,193</sup>, in 10 mL of DM25 in 50 mL Erlenmeyer flasks shaken at 120 rpm over a diameter of 1 inch with incubation at 37°C. After 24 h of growth separately to precondition strains to these conditions, two replicate cultures for each Ara<sup>-</sup> and Ara<sup>+</sup> pair were mixed at equal volumes in fresh DM25 media such that there was an overall 1:100 dilution. Dilutions of these initial mixtures were plated on TA plates to determine the initial representation of each strain in each replicate flask. Then, the competition was carried out over three days of transferring 1:100 dilutions into fresh medium each day. A dilution of each culture after growth on day three was again plated to determine the final representation of each strain. Relative fitness was calculated as the ratio of the realized growth rates of each strain between the final and initial platings<sup>164,192</sup>.

For comparisons of the effect of the authentic *rnk-citG* duplication versus the addition of the *P<sub>rnk</sub>-citT* module to REL606 and ZDB706 (**Figure A1**) we first established

neutrality of an Ara<sup>+</sup> revertant and then judged whether there was significant difference between the fitnesses of the Cit<sup>-</sup> and Cit<sup>+</sup> strains pairs. For comparing the fitness impact of evolving Cit<sup>+</sup> in other strains (**Figure A2**), we measured the relative fitness of the Ara<sup>-</sup> Cit<sup>+</sup> variant of the strain with the P<sub>rnk-citT</sub> module added versus the Ara<sup>+</sup> Cit<sup>-</sup> revertant of its Cit<sup>-</sup> progenitor (Cit competition) and multiplied this by the relative fitness of the Ara<sup>+</sup> Cit<sup>-</sup> revertant versus the Ara<sup>-</sup> Cit<sup>-</sup> clone with the null module added (Ara competition). To account for how error in each of these two competitions impacts confidence in the overall fitness change inferred for evolving Cit<sup>+</sup>, we performed 10,000 bootstrap resamplings of the Ara and Cit competition replicates to estimate 95% fitness intervals and significance on the combined measurements. The same bootstrapping procedure was used for comparing the fitnesses of different strains in the population phylogeny in the procedure that combined them into equivalence groups along the lineage to Cit<sup>+</sup> (**Figure A4**).

#### **qRT-PCR measurement of *citT* expression**

Cells were cultured according to the method described in Blount *et al.*<sup>168</sup>. Briefly, cells were initially grown to saturation in a 5 ml LB culture and transferred into 10 ml of DM25 media (1:10,000 dilution) followed by two 24-hour preconditioning cycles in DM25 with 1:100 dilutions. For each preconditioning cycle, cells were diluted by 1:100 into fresh DM25 media. At this point, we performed a final dilution of 1:100 into DM25. Cells were grown until they reached ~50% of the final OD<sub>420</sub> and the entire culture (10 ml) was harvested for extracting RNA. RNA was extracted from frozen cell pellets using the RNASnap protocol<sup>194</sup>. The resulting supernatant was column purified, incorporating on-column DNase treatment (RNA Clean & Concentrator-25, Zymo Research). TapeStation analysis (Agilent) was used to verify RNA integrity (all

RIN scores  $\geq 8.0$ ). Samples were then reverse transcribed in parallel using random primers, with 200ng of RNA as template (High Capacity Reverse Transcription Kit, Applied Biosystems).

qPCR was run in 384 well plates on an Applied Biosystems ViiA 7, using SYBR Green (Thermo Fisher) as fluorophore in a 5  $\mu$ l reaction. QuantStudio was used to determine quantification cycle (Cq) values. All samples were run in technical triplicates. We selected two reference genes (*refs*), 16S RNA and *idnT*, from an initial pool of candidates based on primer efficiency, primer specificity (as judged by melt curve) and stability of expression in a subset of our strains of interest. Primer efficiency was calculated from the slope of a plot of log(dilution) versus Cq, using a 5-fold or 10-fold dilution series of a pool of cDNA from every sample. Final primer sequences and efficiencies were as follows: *citT* (forward = GTTATAGCGGGTAATGTCTTTC, reverse = CACTGATTGGCCTTGTATTG, efficiency = 99.25%); *idnT* (forward = CCCGACACCGCTATCTACTAATAC, reverse = CGCACCATCGAGCAAATCAT, efficiency = 100.5%); 16S (forward = CCCGAAGGTTAAGCTACCTACT, reverse = CATGAAGTCGGAATCGCTAGTAATC, efficiency = 97.6%). In our final analysis comparing *citT* expression across strains, we used three biological replicates per strain, and 2  $\mu$ l of a 1:100 dilution of cDNA as template. Relative expression ( $R$ ) of *citT* in the strain of interest relative to ancestral REL606 was calculated as follows. First,  $\Delta Cq$  was calculated for individual biological replicates according to  $\Delta Cq = Cq^{citT} - \bar{x}(Cq^{refs})$ , where  $\bar{x}(X)$  represents the mean of the values for quantity  $X$ . Then,  $\Delta\Delta Cq$  and  $R$  were calculated from the mean  $\Delta Cq$  of three biological replicates for each strain tested as  $\Delta\Delta Cq = \bar{x}(\Delta Cq^{strain}) - \bar{x}(\Delta Cq^{REL606})$  and  $R = 2^{-\Delta\Delta Cq}$ .

## Genome sequencing and phylogenetic tree construction

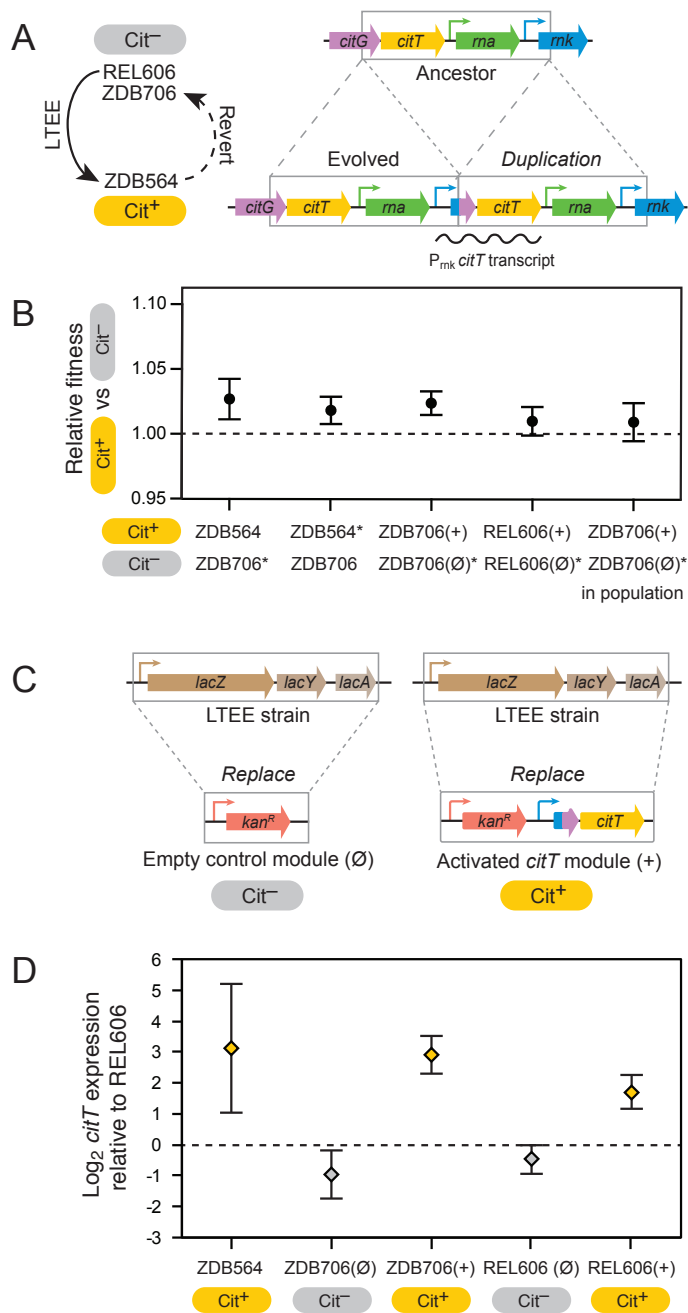
Genome sequences were analyzed for 61 evolved strains from the LTEE population in this study (**Table A1**). For the 20 newly sequenced strains, genomic DNA was purified using the GenElute Bacterial Genomic DNA kit (Sigma) and then sequenced using standard procedures on an Illumina HiSeq 2500 instrument to generate 101-base paired-end reads by the University of Texas at Austin Genome Sequencing and Analysis Facility. Data files for these 20 genomes have been deposited in the NCBI Sequence Read Archive (SRP120037). Raw sequencing reads for all 61 genomes are available via links from the main LTEE NCBI BioProject page (PRJNA414462).

We initially predicted mutations in each re-sequenced genome by comparing Illumina reads to the REL606 reference genome [37] using *breseq* (v0.31.1)<sup>108,195</sup>. Then, we further curated the lists of predicted mutations as previously described<sup>83</sup>. Briefly, a maximum-parsimony phylogenetic tree for all 61 strains from the LTEE population was constructed using the DNAPARS program from the PHYLIP package (v3.69)<sup>196</sup>. Where necessary, we manually corrected mutation predictions, including adding mutations that were hidden by later deletions or splitting sequence differences into multiple mutational events to construct the most parsimonious phylogeny possible. In the current study, we did not discard mutations in repetitive regions before analysis, except we did ignore changes in the hypervariable 7×CCAG repeat at reference coordinates 2103891-2103918 in the final lists of mutations predicted in all clones.

## Beneficial mutation fitness effect models

To construct the curve for the Wiser et al. model<sup>172</sup> in **Figure A4** we calculated the expected time in generations ( $t$ ) and fitness increase ( $s$ ) for each subsequent sweep of a

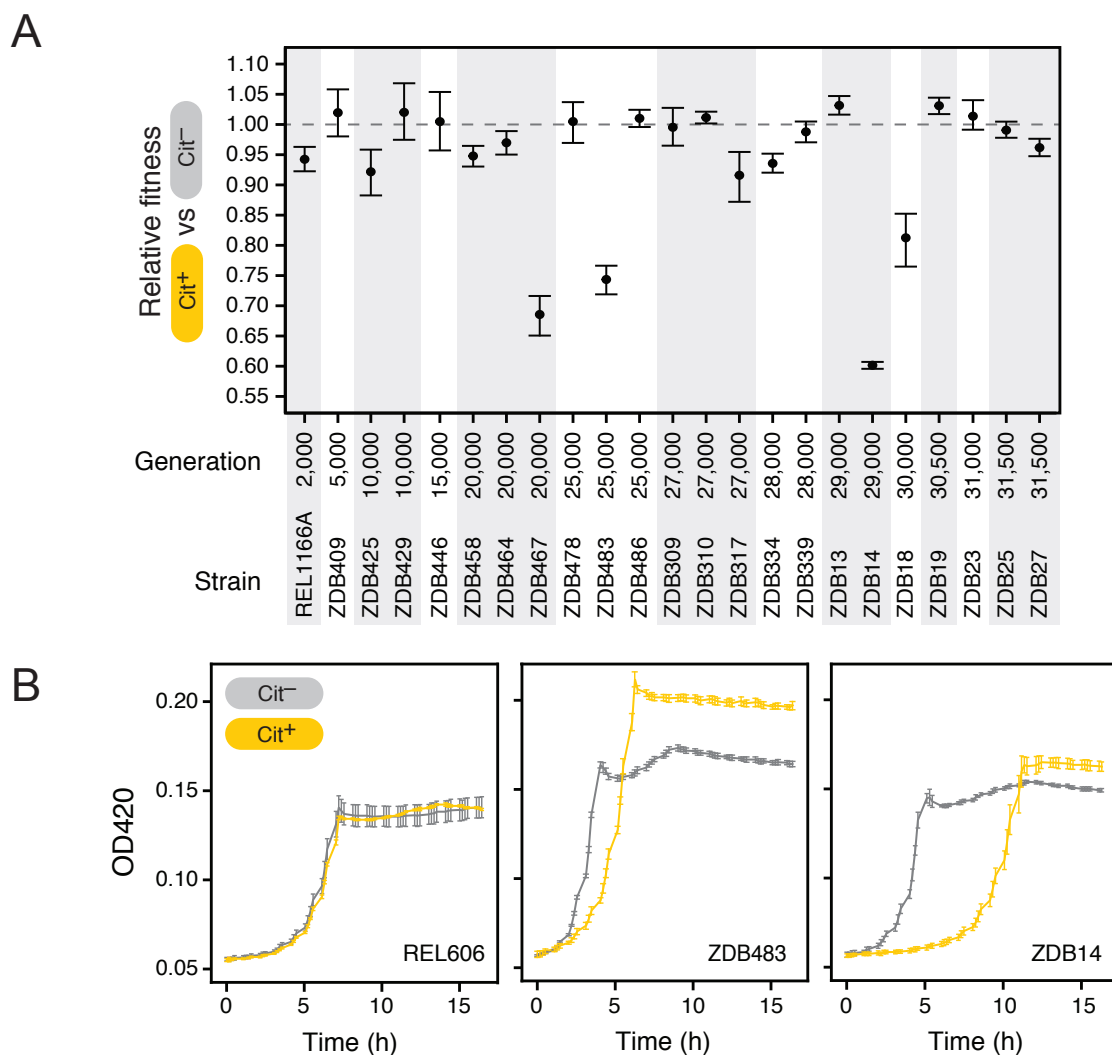
cohort of beneficial mutations using equations S3, S4, and S7 from the supplement of that study using parameter values ( $\alpha_0 = 58.4$ ,  $\mu = 10^{-7}$ , and  $N = 3.3 \times 10^7$ ) that they found to be compatible with the fitness trajectories of the non-mutator LTEE populations. For the Tenaillon et al. model<sup>83</sup>, we first calculated a curve describing the number of beneficial mutations expected in an evolved isolate ( $n$ ) according to the term,  $n = c\sqrt{t}$ , with the best-fit coefficient value ( $c = 0.135$ ) found in that study for all non-mutator LTEE populations considered together. Next, we combined this model with the fitness ( $W$ ) model from Wiser et al.,  $W(t) = (at + 1)^b$ , with best-fit parameters ( $a = 0.0842$  and  $b = 0.00611$ ) found specifically for the citrate population (Ara-3)<sup>172</sup>. Finally, the Tenaillon *et al.*<sup>83</sup> curve in **Figure A4** was graphed by calculating each generation ( $t_n$ ) at which the number of beneficial mutations ( $n$ ) was an integral value,  $t_n = (n/c)^2$ . The graphed selection coefficient was estimated as the fitness at that time  $W(t_n)$  minus the fitness at the time of the previous beneficial mutation  $W(t_{n-1})$ .



**Figure A1. Evolution of rudimentary citrate utilization by activating *citT* expression is slightly beneficial in the genetic background in which it evolved and in the LTEE ancestor**

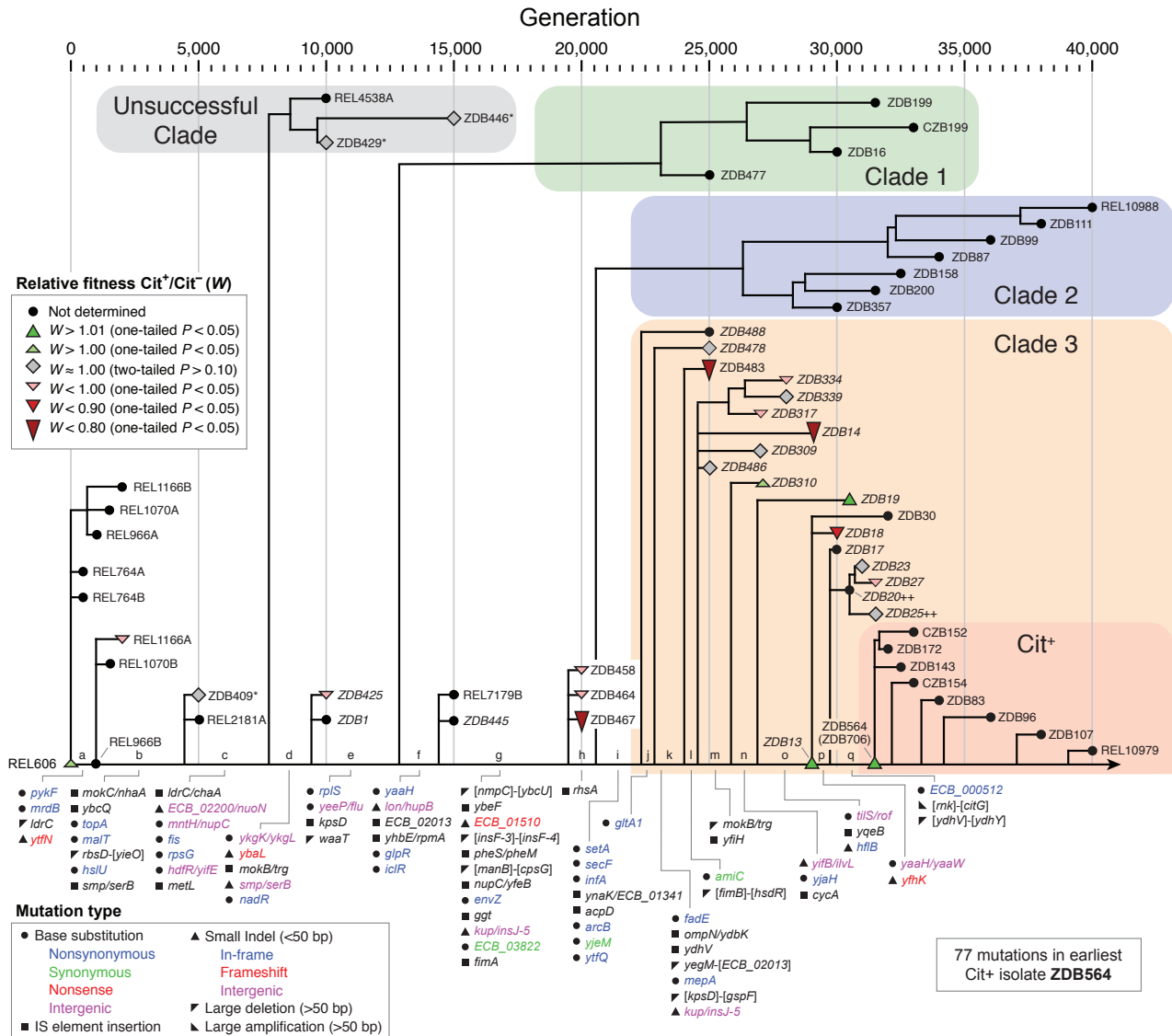


*Figure A1 continued* **(A)** The *rnk-citG* duplication that evolved in the LTEE creates a genomic configuration in which a novel mRNA encoding the CitT transporter is expressed from the *rnk* promoter ( $P_{rnk}$ ) (right). This mutation alone is sufficient for weak citrate utilization (Cit<sup>+</sup> phenotype). It is the ‘actualizing mutation’ in the evolution of this key innovation. Strain ZDB564 is the earliest Cit<sup>+</sup> isolate from the LTEE. In order to measure the effect that this mutation had on competitive fitness when it evolved, a spontaneous Cit<sup>−</sup> revertant of ZDB564 in which the duplication collapsed back to the ancestral state was isolated (left). **(B)** Competitive fitness of Cit<sup>+</sup> versus Cit<sup>−</sup> strain variants. The ZDB564 versus ZDB706 competitions measure the fitness effect of the *rnk-citG* duplication when it evolved. The ZDB706 and REL606 competitions test the effect of adding one copy of the evolved  $P_{rnk}$ -*citT* module into a strain (+) versus adding an empty version of the same cassette (Ø), as pictured in C. An additional ZDB706 competition (in population) was conducted with the two strains together mixed at a 1:99 ratio with the evolved LTEE population from at 31,000 generations to determine if the mutation had a different effect on fitness when rare in the population. Starred strains (\*) have a change to the Ara<sup>+</sup> marker state to allow competition with the corresponding Ara<sup>−</sup> strain. The marker change had no effect on competitive fitness in each case. Error bars are 95% confidence intervals. **(C)** Schematic of the gene cassettes used in the  $P_{rnk}$ -*citT* knock-in assay showing how they were integrated into the *E. coli* chromosome in a way that replaces the native *lac* locus. **(D)** *citT* mRNA expression levels measured relative to the REL606 LTEE ancestor in the evolved Cit<sup>+</sup> isolate from the LTEE (ZDB564) and strains with the  $P_{rnk}$ -*citT* and corresponding empty control cassettes integrated into their chromosomes. Error bars are 95% confidence intervals.



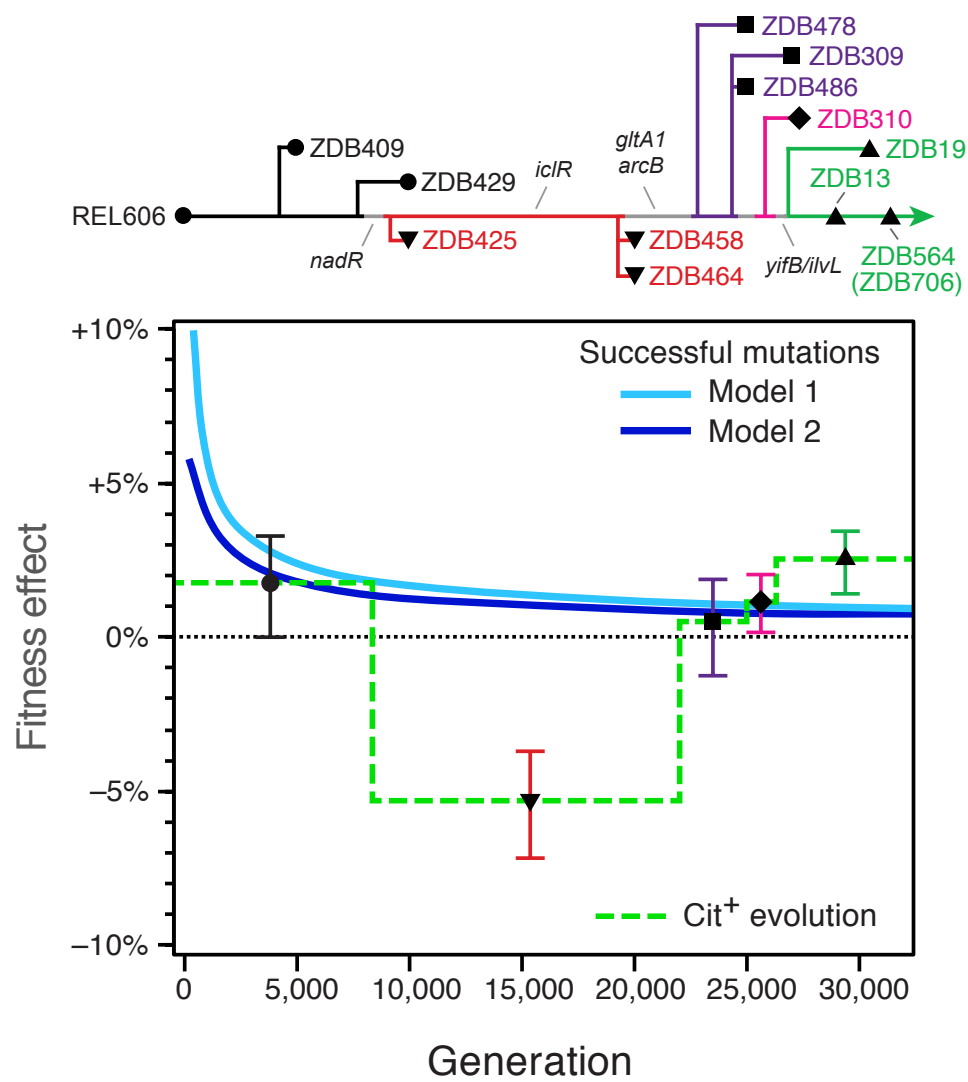
**Figure A2. Fitness consequences of evolving Cit<sup>+</sup> in different evolved genetic backgrounds**

(A) Results of the  $P_{rnk-citT}$  knock-in assay on 23 pre-Cit<sup>+</sup> evolved strains. The clones are ordered by the generation from which they were isolated. Error bars are 95% confidence intervals. Strain construction details and how the results of competition assays were combined into these fitness estimates are described in the **Methods**. (B) Increased lag phase upon addition of the  $P_{rnk-citT}$  module in anti-potentiated strains. Growth curves for the ancestor, REL606, and two anti-potentiated strains, ZDB483 and ZDB14, are shown. Error bars are standard deviations of four replicate cultures.



**Figure A3. Potential for evolving Cit<sup>+</sup> mapped onto phylogeny**

*Figure A3 continued* Phylogeny of isolates from the LTEE population including 20 new clones sequenced for this study to provide better resolution of the timing of mutations on the lineage leading to *Cit*<sup>+</sup> (names in italics). In order to identify changes in the degree of potentiation due to mutations, we mapped the results of the *P<sub>rnk</sub>-citT* knock-in assay onto this phylogenetic tree. Colored symbols reflect the *Cit*<sup>+</sup> to *Cit*<sup>-</sup> relative fitness measured for those strains. The ancestor and 61 evolved isolates were used to construct this phylogenetic tree (**Table A1**). Two clones isolated at 50,000 generations are not shown. Two strains that evolved citrate utilization in replay experiments under the LTEE conditions in a previous study<sup>167</sup> are marked with plus signs (++), and three strains that had evolved alleles added or removed during strain construction are starred (\*).



**Figure A4. Changes in the potential for innovation along the lineage leading to Cit<sup>+</sup> due to genetic and population factors**

*Figure A4 continued* We clustered phylogenetically adjacent strains in which activating *citT* expression had a similar effect on fitness to reconstruct when major changes in potentiation occurred due to new mutations accumulating in the evolved strain (the genetic background). Each cluster is represented by a different color and symbol in the simplified phylogenetic tree (upper panel) and the graph showing the group-wise fitness estimate over a time period (horizontal lines) in which these strains were representative of the main pre-Cit<sup>+</sup> lineage (lower panel). Error bars are 95% confidence intervals. Two models of the rate of adaptation of the LTEE populations at different generations are superimposed on the lower panel. Model 1 estimates the fitness effects of winning cohorts of beneficial mutations sequentially sweeping through the population at each generation according to modelling from Wiser *et al.*<sup>172</sup>. Model 2 estimates the fitness effects of each consecutive beneficial mutation accrued by the winning lineage over time using additional information from Tenaillon *et al.*<sup>83</sup> (see **Methods**). These curves represent competing beneficial mutations that can suppress the evolution of Cit<sup>+</sup> (the population context). If the fitness effect of activating *citT* is above these curves, as it is by 29,000 generations, then it is predicted to be among the most beneficial new mutations that could appear at that time in the LTEE. This means that rudimentary citrate utilization (Cit<sup>+</sup>) can persist in the population long enough to be refined by further mutations to full citrate utilization (Cit<sup>++</sup>); the metabolic innovation can be achieved.

Strain	Generation	Clade	Reference
REL764A	500		Tenaillon et al. 2016
REL764B	500		Tenaillon et al. 2016
REL966A	1,000		Tenaillon et al. 2016
REL966B	1,000		Tenaillon et al. 2016
REL1070A	1,500		Tenaillon et al. 2016
REL1070B	1,500		Tenaillon et al. 2016
REL1166A	2,000		Blount et al. 2012
REL1166B	2,000		Tenaillon et al. 2016
REL2181A	5,000		Tenaillon et al. 2016
ZDB409	5,000		Blount et al. 2012
REL4538A	10,000	UC	Tenaillon et al. 2016
ZDB1	10,000		This study
ZDB425	10,000		This study
ZDB429	10,000	UC	Blount et al. 2012
REL7179B	15,000		Tenaillon et al. 2016
ZDB445	15,000		This study
ZDB446	15,000	UC	Blount et al. 2012
ZDB458	20,000		Blount et al. 2012
ZDB464	20,000		Blount et al. 2012
ZDB467	20,000		Blount et al. 2012
ZDB477	25,000	C1	Blount et al. 2012
ZDB478	25,000	C3	This study
ZDB483	25,000	C3	Blount et al. 2012
ZDB486	25,000	C3	This study
ZDB488	25,000	C3	This study
ZDB309	27,000	C3	This study
ZDB310	27,000	C3	This study
ZDB317	27,000	C3	This study
ZDB334	28,000	C3	This study
ZDB339	28,000	C3	This study
ZDB13	29,000	C3	This study

Strain	Generation	Clade	Reference
ZDB14	29,000	C3	This study
ZDB16	30,000	C1	Blount et al. 2012
ZDB17	30,000	C3	This study
ZDB18	30,000	C3	This study
ZDB357	30,000	C2	Blount et al. 2012
ZDB19	30,500	C3	This study
ZDB20	30,500	C3	This study
ZDB23	31,000	C3	This study
ZDB25	31,500	C3	This study
ZDB27	31,500	C3	This study
ZDB199	31,500	C1	Blount et al. 2012
ZDB200	31,500	C2	Blount et al. 2012
ZDB564	31,500	C3+	Blount et al. 2012
ZDB30	32,000	C3+	Blount et al. 2012
ZDB172	32,000	C3+	Blount et al. 2012
ZDB143	32,500	C2	Blount et al. 2012
ZDB158	32,500	C2	Blount et al. 2012
CZB152	33,000	C3+	Blount et al. 2012
CZB154	33,000	C3+	Blount et al. 2012
CZB199	33,000	C1	Blount et al. 2012
ZDB83	34,000	C3+	Blount et al. 2012
ZDB87	34,000	C2	Blount et al. 2012
ZDB96	36,000	C3+H	Blount et al. 2012
ZDB99	36,000	C2	Blount et al. 2012
ZDB107	38,000	C3+H	Blount et al. 2012
ZDB111	38,000	C2	Blount et al. 2012
REL10979	40,000	C3+H	Blount et al. 2012
REL10988	40,000	C2	Blount et al. 2012
REL11364	50,000	C3+H	Tenaillon et al. 2016
REL11365	50,000	C3+H	Tenaillon et al. 2016

**Table A1. Genome sequencing of *E. coli* isolates from the LTEE population**

Clade designations describe placement in the phylogenetic tree of all sequenced strains from the Ara-3 population and relative to key evolutionary transitions in this population: UC, Unsuccessful Clade; C1, Clade 1; C2, Clade 2; C3, Clade 3; C3+, Clade 3 Cit+; C3+H, Clade 3 Cit+ hypermutator.

Strain	Generation	Description <sup>†</sup>
DL82	0	REL606(+)
DL86	0	REL607(Ø)*
DL426	2,000	REL1166A*
EQ966	2,000	REL1166A(+)
DL631	2,000	REL1166A(Ø)*
DL403	5,000	ZDB409*
DL479	5,000	ZDB409(+)
DL945	5,000	ZDB409(Ø)*
DL418	10,000	ZDB425*
DL439	10,000	ZDB425(+)
DL959	10,000	ZDB425(Ø)*
DL424	10,000	ZDB429*
EQ972	10,000	ZDB429(+)
DL480	10,000	ZDB429(Ø)*
DL406	15,000	ZDB446*
DL475	15,000	ZDB446(+)
DL633	15,000	ZDB446(Ø)*
DL415	20,000	ZDB458*
DL13	20,000	ZDB458(+)
DL376	20,000	ZDB458(Ø)*
DL368	20,000	ZDB464*
DL11	20,000	ZDB464(+)
DL379	20,000	ZDB464(Ø)*
DL366	20,000	ZDB467*
DL12	20,000	ZDB467(+)
DL377	20,000	ZDB467(Ø)*
DL76	25,000	ZDB478*
DL84	25,000	ZDB478(+)
DL93	25,000	ZDB478(Ø)*
DL363	25,000	ZDB483*
DL15	25,000	ZDB483(+)
DL374	25,000	ZDB483(Ø)*
DL129	25,000	ZDB486*
DL433	25,000	ZDB486(+)
DL612	25,000	ZDB486(Ø)*
DL137	27,000	ZDB309*

Strain	Generation	Description <sup>†</sup>
DL158	27,000	ZDB309(+)
DL159	27,000	ZDB309(Ø)*
DL412	27,000	ZDB310*
DL463	27,000	ZDB310(+)
DL533	27,000	ZDB310(Ø)*
DL409	27,000	ZDB317*
DL471	27,000	ZDB317(+)
DL957	27,000	ZDB317(Ø)*
DL421	28,000	ZDB334*
DL467	28,000	ZDB334(+)
DL485	28,000	ZDB334(Ø)*
DL359	28,000	ZDB339*
EQ1104	28,000	ZDB339(+)
DL370	28,000	ZDB339(Ø)*
DL161	29,000	ZDB13*
DL617	29,000	ZDB13(+)
DL175	29,000	ZDB13(Ø)*
DL361	29,000	ZDB14*
EQ1068	29,000	ZDB14(+)
DL372	29,000	ZDB14(Ø)*
DL163	30,000	ZDB18*
EQ1111	30,000	ZDB18(+)
DL176	30,000	ZDB18(Ø)*
DL164	30,500	ZDB19*
EQ1113	30,500	ZDB19(+)
DL180	30,500	ZDB19(Ø)*
DL266	31,000	ZDB23*
DL310	31,000	ZDB23(+)
DL308	31,000	ZDB23(Ø)*
DL167	31,500	ZDB25*
DL201	31,500	ZDB25(+)
DL185	31,500	ZDB25(Ø)*
DL261	31,500	ZDB27*
DL314	31,500	ZDB27(+)
DL312	31,500	ZDB27(Ø)*

**Table A2. Strains constructed in this study**

<sup>†</sup>Symbol legend: \*for Ara<sup>+</sup> revertant, (+) for P<sub>rnk-citT</sub> cassette, (Ø) for empty control cassette



## **Appendix B: Bacterial production of gellan gum as a do-it-yourself alternative to agar**

**This chapter is reproduced (with minor modifications) from its initial publication:**

McGuffey JC‡ and **Leon D‡**, Dhanji EZ, Mishler DM, Barrick JE. (2018) Bacterial production of gellan gum as a do-it-yourself alternative to agar. *J Microbiol Biol Educ* **19**,182-184.

‡ designates equal contribution

### **Research Contributions**

Leon D, Mishler DM, and Barrick JE designed experiments. Leon D, McGuffey JC, and Dhanji EZ performed experiments. Leon D, McGuffey JC, and Barrick JE analyzed data. Leon D and Barrick JE wrote the bulk of the manuscript.

### **Acknowledgements**

We thank the 2016 UT Austin iGEM team and Microbe Hackers Freshman Research Initiative stream for helpful discussions. We acknowledge support from the National Science Foundation (NSF) (CBET-1554179) and the NSF BEACON Center for the Study of Evolution in Action (DBI-0939454) for work on this protocol. J.M.C. acknowledges a UT Austin TIDES fellowship. D.L. acknowledges an ASM Watkins Graduate Research Fellowship. The authors declare that there are no conflicts of interest.

## ABSTRACT

Lack of access to expensive reagents and equipment are barriers to performing microbiology experiments in K-12 classrooms and do-it-yourself (DIY) science settings. We describe a procedure for using the bacterium *Sphingomonas paucimobilis* to synthesize gellan gum as an affordable alternative to purchasing agar for educators and DIY scientists. The method involves microwaving gelatinous cultures of *S. paucimobilis* ATCC 31461 after a two-step growth procedure, and then pouring plates. Gellan gum produced by *S. paucimobilis* acts as a solidifying agent and provides a resilient surface that supports the growth of microbial colonies. This DIY procedure offers an opportunity to experiment with microbial production of an extracellular polysaccharide and to cheaply and sustainably source a reagent for research.

## INTRODUCTION

Cost can be a steep barrier to participation in science. As a result, most research takes place in universities, government, and industry. Recently, a do-it-yourself (DIY) biology community has emerged outside of this system<sup>197</sup>. DIY scientists often conduct their experiments in non-traditional settings (e.g., kitchens, garages, maker spaces) and fund their own work. To make microbiology more affordable in this and other educational environments, we became interested in a DIY replacement for agar.

Agar is one of the most expensive and routinely utilized reagents in microbiology labs. It is used as a gelling agent in solid media for growing colonies. Agar is a mixture of polysaccharides that is processed from the cell walls of seaweed (red algae). These algae are currently harvested from the wild, which has contributed to fluctuations and shortages in the agar supply, most recently in 2015<sup>198</sup>. Microbes are already the source of many nutrients (e.g., yeast extract) and supplements (e.g., amino acids and antibiotics) used in common culture media. What if microbes could also cheaply produce a gelling agent?

They can! Gellan gum is a capsular exopolysaccharide (EPS) that is naturally synthesized by the bacterium *Sphingomonas paucimobilis* ATCC 31461<sup>199,200</sup>. Gellan gum is widely used as a solidifying agent in the food industry, and it is sold as a chemically purified product (e.g., Gelrite®) for use in microbial growth media<sup>199</sup>. Purification of gellan gum normally involves cultivating *S. paucimobilis*, several precipitation steps to separate the polysaccharide from cells, and a final drying phase<sup>201</sup>. Here, we present a streamlined procedure for affordably producing gellan gum plates that requires only DIY equipment and ingredients.

## METHODS AND RESULTS

### I. Prepare DIY media for gellan gum production

Culturing *S. paucimobilis* for gellan gum production is a two-step process in which a starter culture is propagated in a rich medium and then transferred to a minimal medium. The initial culture phase allows cells to rapidly achieve high densities before they are used to

inoculate gellan production medium, where little further growth occurs<sup>202</sup>. Gellan gum production is induced under nitrogen starvation in the presence of excess carbon, which serves as the substrate for EPS synthesis. Therefore, gellan production media contains a high C:N ratio to favor maximal EPS accumulation<sup>199</sup>.

When carrying out this procedure, you should wear appropriate personal protective equipment such as safety goggles, gloves, and a lab coat and use heat resistant glassware (e.g., Pyrex). For the two-step culturing method, you will need to first make a DIY rich medium, General Kitchen Broth (GKB)<sup>203</sup>, which substitutes for a conventional *S. paucimobilis* rich medium used in the first step (e.g., YPG). To make GKB, mix the following ingredients in 100 mL of water: 0.5 g dried skim milk, 0.25 g marmite, and 0.1 g honey. For the second step, we developed a minimal medium recipe (DIY-GPM) by identifying DIY equivalents for components of standard gellan production medium<sup>202</sup>. To make DIY-GPM, mix the following in 1 L of water: 0.03 g dried skim milk, 0.5 g marmite, 24.35 g honey, 1 g table salt, 1.2 g Epsom salt, 10 g trisodium phosphate, 3.75 mL clear ammonia (e.g., Austin's brand, ~2.5% w/v ammonium hydroxide) and 3.1 g citric acid. Ensure that the GKB and DIY-GPM components are well mixed and sterilize them, using an autoclave or pressure cooker<sup>203</sup>.

## II. Culture *S. paucimobilis* for gellan gum production

Once the media has cooled to room temperature, start a 100 mL culture of *Sphingomonas paucimobilis* ATCC 31461 in GKB. Allow this to grow to saturation by incubating for 2 days, with shaking and at 30°C if possible (**Figure B1A**). Next, add this entire culture to 1 L of sterile DIY-GPM and incubate under the same conditions until the culture becomes gelatinous and homogenous. We found that incubating for eight days at 30°C with shaking produced a sufficient yield of gellan gum. You should inspect these cultures every day. A culture that is ready for pouring plates should be highly viscous and adhere to the side of the flask when it is tilted (**Figure B1B**).

### III. Pour gellan gum plates

Next, add 0.4 g/L more of Epsom salt to the *S. paucimobilis* culture. This provides magnesium to strengthen the gellan polymer matrix<sup>199</sup>. Also, add any other nutrients needed by your microbe of interest<sup>203</sup>. For example, we found that adding 10 g/L of marmite improved growth of *E. coli* and *S. cerevisiae*. Heat the entire culture in a microwave using a medium setting until it becomes fluid and homogenous, watching carefully to avoid overboiling. At this point, the mixture is sterile and can be poured into petri dishes to make ~40 standard plates. We observed normal colony growth rates and morphology on these DIY gellan gum plates (**Figure B1C**).

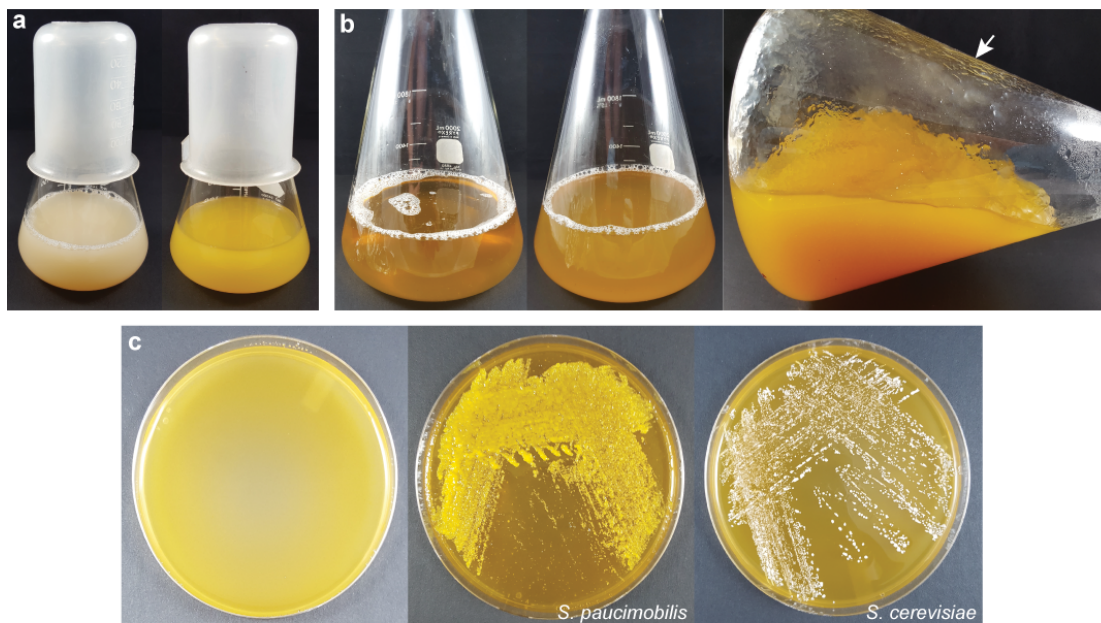
### Safety issues

*S. paucimobilis* ATCC 31461 is classified as a biosafety level 1 (BSL1) organism. Students and DIY scientists should follow the American Society for Microbiology's guidelines for teaching laboratories which describe appropriate personal protective equipment and sterile procedures for working safely with BSL1 microbes<sup>204</sup>.

### DISCUSSION

This protocol can be used sustainably and affordably source a homegrown agar replacement. We estimate that the cost of the culture components needed for DIY gellan gum production is roughly one tenth the cost of agar (\$0.58 versus \$5.60 per liter of media). Cultures of *S. paucimobilis* become extremely gelatinous, which also makes it an interesting microbe for classroom demonstrations related to the materials properties of extracellular polysaccharides. *S. paucimobilis* strains used for commercial production have mutations to make them unpigmented or to produce deacetylated variants of gellan gum with improved gelling properties<sup>205,206</sup>. It is possible that DIY researchers or students could also alter these properties or even engineer *S. paucimobilis* variants that incorporate

metabolic pathways for manufacturing other expensive media components (e.g., cofactors or antibiotics) in the future.



**Figure B1. DIY gellan gum plates**

(A) Initial phase of culturing *S. paucimobilis* in rich media. GKB medium before inoculation with *S. paucimobilis* (left) and after 2 days of growth (right). (B) Second stage of culturing for gellan gum production. DIY-GPM medium before inoculation (left), immediately after inoculation (middle), and after 8 days of growth (right). The arrow points to gellan gum adhering to the side of the flask. (C) DIY gellan gum plates supplemented with marmite support the growth of colonies of *S. paucimobilis*, *S. cerevisiae*, and *E. coli* (not shown).

## REFERENCES

1. Endy D. Foundations for engineering biology. (2005) *Nature* **438**, 449–453.
2. Arkin AP. (2013) A wise consistency: engineering biology for conformity, reliability, predictability. *Curr Opin Chem Biol* **17**, 893–901.
3. Venturelli OS, Egbert RG and Arkin AP. (2016) Towards Engineering Biological Systems in a Broader Context. *J Mol Biol* **428**, 928–944.
4. Brophy JAN and Voigt CA. (2014) Principles of genetic circuit design. *Nat Meth* **11**, 508–520.
5. Canton B, Labno A and Endy D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat Biotechnol* **26**, 787–793.
6. Kwok, R. (2010) Five hard truths for synthetic biology. *Nature* **463**, 288–290.
7. Boyer H. (1978) First Successful Laboratory Production of Human Insulin Announced. *Genentech News Release* 9/6/1978.
8. Paddon CJ and Keasling JD. (2014) Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat Rev Micro* **12**, 355–367.
9. Peplow M. (2016) Synthetic malaria drug meets market resistance. *Nature* **530**, 389–390.
10. Teng, X *et al.* (2013) Genome-wide consequences of deleting any single gene. *Mol Cell* **52**, 485–494.
11. Borkowski O, Ceroni F, Stan GB and Ellis T. (2016) Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. *Curr Opin Microbiol* **33**, 123–130.
12. Bentley WE *et al.* (1989) Plasmid-encoded protein: The principal factor in the ‘metabolic burden’ associated with recombinant bacteria. *Biotechnol Bioeng* **102**, 1283–1297.
13. Liu Q, Schumacher J, Wan X, Lou C and Wang B. (2018) Orthogonality and Burdens of Heterologous AND Gate Gene Circuits in *E. coli*. *ACS Synth Biol* **7**, 553–564.
14. Carrera J, Rodrigo G, Singh V, Kirov B and Jaramillo A. (2011) Empirical model and *in vivo* characterization of the bacterial response to synthetic gene expression show that ribosome allocation limits growth rate. *Biotechnol J* **6**, 773–783.
15. Cardinale S and Arkin AP. (2012) Contextualizing context for synthetic biology – identifying causes of failure of synthetic biological systems. *Biotechnol J* **7**, 856–866.
16. Arkin, A P and Fletcher DA. (2006) Fast, cheap and somewhat in control. *Genome Biol* **7**, 114.
17. Dong H, Nilsson L and Kurland CG. (1995) Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *J Bacteriol* **177**, 1497–1504.



18. Vind J, Sørensen MA, Rasmussen MD and Pedersen S. (1993) Synthesis of proteins in *Escherichia coli* is limited by the concentration of free ribosomes. Expression from reporter genes does not always reflect functional mRNA levels. *Biochem Mol Biol* **231**, 678–688.
19. Ow DS, Nissom PM, Philp R, Oh SK and Yap MGS. (2006) Global transcriptional analysis of metabolic burden due to plasmid maintenance in *Escherichia coli* DH5 $\alpha$  during batch fermentation. *Enzyme Microb Technol* **39**, 391–398.
20. Flores S, de Anda-Herrera R, Gosset G and Bolívar FG. (2004) Growth-rate recovery of *Escherichia coli* cultures carrying a multicopy plasmid, by engineering of the pentose-phosphate pathway. *Biotechnol Bioeng* **87**, 485–494.
21. Neubauer P, Lin HY and Mathisizik B. (2003) Metabolic load of recombinant protein production: inhibition of cellular capacities for glucose uptake and respiration after induction of a heterologous gene in *Escherichia coli*. *Biotechnol Bioeng* **83**, 53–64.
22. Sleight SC, Bartley BA, Lieviant JA and Sauro HM. (2010) Designing and engineering evolutionary robust genetic circuits. *J Biol Eng* **4**, 12.
23. Rugbjerg P, Myling-Petersen N, Porse A, Sarup-Lytzen K and Sommer MOA. (2018) Diverse genetic error modes constrain large-scale bio-based production. *Nat Commun* **9**, 787.
24. You L, Cox RS, Weiss R and Arnold FH. (2004) Programmed population control by cell-cell communication and regulated killing. *Nature* **428**, 868–871.
25. Darmon E and Leach DRF. (2014) Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* **78**, 1–39.
26. Schroeder SW, Ponlkrity Y, Simmons LA and Wang JD. (2017) Sources of spontaneous mutagenesis in bacteria. *Crit Rev Biochem Mol Biol* **53**, 29–48.
27. Maki H. (2002) Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annu Rev Genet* **36**, 279–303.
28. Friedberg EC, Walker GC, Siede W, Wood, RD, Schultz RA, and Ellenberger T. (2006) DNA repair pathways and mechanisms 2<sup>nd</sup> Ed. *ASM Press*, Washington D.C.
29. Kim N and Jinks-Robertson S. (2012) Transcription as a source of genome instability. *Nat Rev Genet* **13**, 204–214.
30. Sollier J and Cimprich KA. (2015) Breaking bad: R-loops and genome integrity. *Trends Cell Biol* **25**, 514–522.
31. Schaaper, RM. (1998) Antimutator mutants in bacteriophage T4 and *Escherichia coli*. *Genetics* **148**, 1579–1585.
32. Foster PL, Lee H, Popodi E, Townes JP, Tang H. (2015) Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proc Natl Acad Sci U S A* **112**:E5990-9.

33. Herr AJ, Williams LN and Preston BD. (2011) Antimutator variants of DNA polymerases. *Crit Rev in Biochem Mol Biol* **46**, 548–570.
34. Drake JW, Allen EF, Forsberg SA, *et al.* (1969) Genetic control of mutation rates in bacteriophage T4. *Nature* **221**, 1128–1132.
35. Oller AR and Schaaper RM. (1994) Spontaneous mutation in *Escherichia coli* containing the *dnaE911* DNA polymerase antimutator allele. *Genetics* **138**, 263–270.
36. Loh E, Choe J and Loeb LA. (2007) Highly tolerated amino acid substitutions increase the fidelity of *Escherichia coli* DNA polymerase I. *J Biol Chem* **282**, 12201–12209.
37. Quiñones A and Piechocki R. (1985) Isolation and characterization of *Escherichia coli* antimutators. A new strategy to study the nature and origin of spontaneous mutations. *Mol Gen Genet* **201**, 315–322.
38. Fijalkowska IJ, Dunn RL and Schaaper RM. (1993) Mutants of *Escherichia coli* with increased fidelity of DNA replication. *Genetics* **134**, 1023–1030.
39. Csörgő B, Fehér T, Tímár E, Blattner FR and Pósfai G. (2012) Low-mutation-rate, reduced-genome *Escherichia coli*: an improved host for faithful maintenance of engineered genetic constructs. *Microb Cell Fact* **11**, 11.
40. Harris RS *et al.* (1997) Mismatch repair protein MutL becomes limiting during stationary-phase mutation. *Genes Dev* **11**, 2426–2437.
41. Kamiya H *et al.* (2003) Suppression of spontaneous and hydrogen peroxide-induced mutations by a MutT-type nucleotide pool sanitization enzyme, the *Escherichia coli* Orf135 protein. *Curr Protoc Mol Biol* **8**, 941-950.
42. Sandoval CM *et al.* (2014) Use of pantothenate as a metabolic switch increases the genetic stability of farnesene producing *Saccharomyces cerevisiae*. *Metab Eng* **25**, 215–226.
43. Gibson DG *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56.
44. Wannier TM *et al.* (2018) Adaptive evolution of genomically recoded *Escherichia coli*. *Proc Natl Acad Sci U S A* **115**, 3090–3095.
45. Bull JJ and Barrick JE. (2017) Arresting Evolution. *Trends Genet* **33**, 910–920.
46. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. (2010) Interdependence of cell growth and gene expression: origins and consequences. *Science* **330**, 1099–1102.
47. Renda BA, Hammerling MJ and Barrick JE. (2014) Engineering reduced evolutionary potential for synthetic biology. *Mol BioSyst* **10**, 1668–1678.
48. Pósfai G *et al.* (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–1046.
49. Jack BR *et al.* (2015) Predicting the genetic stability of engineered DNA sequences with the EFM calculator. *ACS Synth Biol* **4**, 939–943.

50. Suárez GA, Renda BA, Dasgupta A and Barrick JE. (2017) Reduced mutation rate and increased transformability of transposon-free *Acinetobacter baylyi* ADP1-ISx. *Appl Environ Microbiol* **83**, 17.
51. Choi JW, Yim SS, Kim MJ and Jeong KJ. (2015) Enhanced production of recombinant proteins with *Corynebacterium glutamicum* by deletion of insertion sequences (IS elements). *Microb Cell Fact* **14**, 207.
52. Ceroni F, Algar R, Stan GB and Ellis T. (2015) Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nat Meth* **12**, 415–418.
53. Romero PA and Arnold FH. (2009) Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**, 1–12.
54. Haseltine EL and Arnold FH. (2007) Synthetic gene circuits: design with directed evolution. *Annu Rev Biophys Biomol Struct* **36**, 1–19.
55. Baba T *et al.* (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 473–11.
56. Wang F and Yang W. (2009) Structural insight into translesion synthesis by DNA Pol II. *Cell* **139**, 1279–1289.
57. Lin-Chao S and Cohen SN. (1991) The rate of processing and degradation of antisense RNAI regulates the replication of ColE1-type plasmids *in vivo*. *Cell* **65**, 1233–1242.
58. Kües U and Stahl U. (1989) Replication of plasmids in gram-negative bacteria. *Microbiol Rev* **53**, 491–516.
59. Tomcsányi T and Apirion D. (1985) Processing enzyme ribonuclease E specifically cleaves RNA I. An inhibitor of primer formation in plasmid DNA synthesis. *J Mol Biol* **185**, 713–720.
60. Callaghan AJ *et al.* (2005) Structure of *Escherichia coli* RNase E catalytic domain and implications for RNA turnover. *Nature* **437**, 1187–1191.
61. Beese LS, Friedman, JM, Steitz TA. (1993) Crystal structures of the Klenow fragment of DNA polymerase I complexed with deoxynucleoside triphosphate and pyrophosphate. *ACS Biochem* **32**, 14095–14101.
62. Rosche WA and Foster PL. (2017) Determining mutation rates in bacterial populations. *Methods* **20**, 1–14.
63. Tack DS *et al.* (2016) Addicting diverse bacteria to a noncanonical amino acid. *Nat Chem Biol* **12**, 138–140.
64. Bataillon T and Bailey SF. (2014) Effects of new mutations on fitness: insights from models and data. *Ann N Y Acad Sci* **1320**, 76–92.
65. Eyre-Walker A and Keightley PD. (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**, 610–618.
66. Robert L *et al.* (2018) Mutation dynamics and fitness effects followed in single cells. *Science* **359**, 1283–1286.

67. Sniegowski PD, Gerrish PJ, Johnson T and Shaver A. (2000) The evolution of mutation rates: separating causes from consequences. *Bioessays* **22**, 1057–1066.
68. de Visser JA. (2002) The fate of microbial mutators. *Micro Soc* **148**, 1247–1252.
69. Denamur E and Matic I. (2006) Evolution of mutation rates in bacteria. *Mol Microbiol* **60**, 820–827.
70. Voordeckers K *et al.* (2015) Adaptation to high ethanol reveals complex evolutionary pathways. *PLoS Genet* **11**, e1005635.
71. Sniegowski PD, Gerrish PJ and Lenski RE. (1997) Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* **387**, 703–705.
72. Tenaillon O, Taddei F, Radmian M and Matic I. (2001) Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res Microbiol* **152**, 11–16.
73. Barrick JE and Lenski RE. (2013) Genome dynamics during experimental evolution. **14**, 827–839.
74. Funchain P *et al.* (2000) The consequences of growth of a mutator strain of *Escherichia coli* as measured by loss of function among multiple gene targets and loss of fitness. *Genetics* **154**, 959–970.
75. Leiby N and Marx CJ. (2014) Metabolic erosion primarily through mutation accumulation, and not tradeoffs, drives limited evolution of substrate specificity in *Escherichia coli*. *PLoS Biol* **12**, e1001789.
76. Marvig RL, Sommer LM, Molin S and Johansen HK. (2015) Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat Genet* **47**, 57–64.
77. Oliver A and Mena A. (2010) Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. *Clin Microbiol Infect* **16**, 798–808.
78. Oliver A, Cantón R, Campo P, Baquero F and Blázquez J. (2000) High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* **288**, 1251–1253.
79. Couce A *et al.* (2017) Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proc Natl Acad Sci U S A* **114**, E9026–E9035.
80. McDonald MJ *et al.* (2012) The evolution of low mutation rates in experimental mutator populations of *Saccharomyces cerevisiae*. *Curr Biol* **22**, 1235–1240.
81. Sprouffske K, Aguilar-Rodríguez J, Sniegowski P and Wagner A. (2018) High mutation rates limit evolutionary adaptation in *Escherichia coli*. *PLoS Genet* **14**, e1007324–31.
82. Wielgoss S *et al.* (2013) Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A* **110**, 222–227.

83. Tenaillon O *et al.* (2016) Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165–170.
84. Lynch M. (2010) Evolution of the mutation rate. *Trends Genet* **26**, 345–352.
85. Lynch M. *et al.* (2016) Genetic drift, selection and the evolution of the mutation rate. *Nature* **17**, 704–714.
86. Friedberg EC, Walker GC, Siede W and Wood RD. (2005) *DNA repair and mutagenesis*.
87. Galperin MY, Moroz OV, Wilson KS and Murzin AG. (2006) House cleaning, a part of good housekeeping. *Mol Microbiol* **59**, 5–19.
88. Fujikawa K and Kasai H. (2002) The oxidized pyrimidine ribonucleotide, 5-hydroxy-CTP, is hydrolyzed efficiently by the *Escherichia coli* recombinant Orf135 protein. *DNA Repair* **1**, 571–576.
89. Loh E, Salk JJ and Loeb LA. (2010) Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc Natl Acad Sci U S A* **107**, 1154–1159.
90. Camps M *et al.* (2003) Targeted gene evolution in *Escherichia coli* using a highly error-prone DNA polymerase I. *Proc Natl Acad Sci U S A* **100**, 9727–9732.
91. Yang YL and Polisky B. (1993) Suppression of ColE1 high-copy-number mutants by mutations in the polA gene of *Escherichia coli*. *J Bacteriol* **175**, 428–437.
92. Frisch RL *et al.* (2010) Separate DNA Pol II- and Pol IV-dependent pathways of stress-induced mutation during double-strand-break repair in *Escherichia coli* are controlled by RpoS. *J Bacteriol* **192**, 4694–4700.
93. Dapa T, Fleurier S, Bredeche MF and Matic I. (2017) The SOS and RpoS regulons contribute to bacterial cell robustness to genotoxic stress by synergistically regulating DNA polymerase Pol II. *Genetics* **206**, 1349–1360.
94. Witkin EM. (1976) Ultraviolet mutagenesis and inducible DNA repair in *Escherichia coli*. *Bacteriol Rev* **40**, 869–907.
95. Janel-Bintz R, *et al.* (2017) Processing closely spaced lesions during Nucleotide Excision Repair triggers mutagenesis in *E. coli*. *PLoS Genet* **13**, 1–27.
96. Escarceller M. *et al.* (1994) Involvement of *Escherichia coli* DNA polymerase II in response to oxidative damage and adaptive mutation. *J Bacteriol* **176**, 6221–6228.
97. Hastings PJ *et al.* (2000) Competition of *Escherichia coli* DNA polymerases I, II and III with DNA Pol IV in stressed cells. *PLoS ONE* **5**, e10862–11.
98. Napolitano R, Janel-Binz R, Wagner J, and Fuchs RPP. (2000) All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis. *EMBO J* **19**, 6259–6265.
99. Wagner J, Etienne H, Janel-Bintz R, Fuchs RP. (2002) Genetics of mutagenesis in *E. coli*: various combinations of translesion polymerases (Pol

- II, IV and V) deal with lesion/sequence context diversity. *DNA Repair* **1**, 159–167.
100. Cirz RT *et al.* (2005) Inhibition of mutation and combating the evolution of antibiotic resistance. *PLoS Biol* **3**, e176–10.
  101. Mohanty BK and Kushner SR. (2016) Regulation of mRNA decay in bacteria. *Annu Rev Microbiol* **70**, 25–44.
  102. Stead MB *et al.* (2010) Analysis of *Escherichia coli* RNase E and RNase III activity *in vivo* using tiling microarrays. *Nucleic Acids Res* **39**, 3188–3203.
  103. Chao Y *et al.* (2017) *In vivo* cleavage map illuminates the central role of RNase E in coding and non-coding RNA pathways. *Mol Cell* **65**, 39–51.
  104. Heller L. (2007) Small RNA regulators and the bacterial response to stress. *Cold Spring Harb Symp Quant Biol* **71**, 1–12.
  105. Wang HH *et al.* (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898.
  106. Garst AD *et al.* (2016) Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nature* **35**, 1–13.
  107. Umenhoffer K *et al.* (2010) Reduced evolvability of *Escherichia coli* MDS42, an IS-less cellular chassis for molecular and synthetic biology applications. *Microb Cell Fact* **9**, 38.
  108. Deatherage DE and Barrick JE. (2014) Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol* **1151**, 165–188 (2014).
  109. Nehring RB. *et al.* (2016) An ultra-dense library resource for rapid deconvolution of mutations that cause phenotypes in *Escherichia coli*. *Nucleic Acids Res* **44**, e41–e41.
  110. Thomason, LC, Costantino N and Court DL. (2007) *E. coli* genome manipulation by P1 transduction. *Curr Protoc Mol Biol* **Chapter 1**, Unit 1.17.
  111. Datsenko KA and Wanner BL. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**, 6640–6645.
  112. Foster, P. L. (2006) Methods for determining spontaneous mutation rates. *Meth Enzymol* **409**, 195–213.
  113. Jin DJ and Zhou YN. (1996) Mutational analysis of structure-function relationship of RNA polymerase in *Escherichia coli*. *Meth Enzymol* **273**, 300–319.
  114. Fehér T, Cseh B, Umenhoffer K, Karcagi I and Pósfai G. (2006) Characterization of *cycA* mutants of *Escherichia coli*. *Mutat Res* **595**, 184–190.
  115. Zheng Q. (2017) rSalvador: An R Package for the Fluctuation Experiment. *G3 (Bethesda)* **7**, 3849–3856.

116. Lee C, Kim J, Shin SG and Hwang S. (2006) Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. *J Biotechnol* **123**, 273–280.
117. Annaluru N *et al.* (2014) Total synthesis of a functional designer eukaryotic chromosome. *Science* **344**, 55–58.
118. Deatherage DE, Leon D, Rodriguez ÁE, Omar S, Barrick JE. (2018) Directed evolution of *Escherichia coli* with lower-than-natural plasmid mutation rates. *Nucleic Acids Res* (in press).
119. Fraser, M. E., James, M. N., Bridger, W. A. & Wolodko, W. T. (1999) A detailed structural description of *Escherichia coli* succinyl-CoA synthetase. *J Mol Biol* **285**, 1633–1653.
120. Buck D, Spencer ME, Guest JR. (1985) Primary structure of the succinyl-CoA synthetase of *Escherichia coli*. *Biochemistry* **22**, 6245–6252.
121. Cecchini G. (2003) Function and structure of complex II of the respiratory chain. *Annu Rev Biochem* **72**, 77–109.
122. Yankovskaya V *et al.* (2003) Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science* **299**, 700–704.
123. Imlay JA. (2013) The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. *Nat Rev Micro* **11**, 443–454.
124. Imlay JA. (2008) Cellular defenses against superoxide and hydrogen peroxide. *Annu Rev Biochem* **77**, 755–776.
125. Imlay JA. (2015) Diagnosing oxidative stress in bacteria: not as easy as you might think. *Curr Opin Microbiol* **24**, 124–131.
126. Sandoval JM, Arenas FA and Vásquez CC. (2011) Glucose-6-phosphate dehydrogenase protects *Escherichia coli* from tellurite-mediated oxidative stress. *PLoS ONE* **6**, e25573.
127. Studier FW and Moffatt BA. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* **189**, 113–130.
128. Studier FW, Rosenberg AH, Dunn JJ and Dubendorff JW. (1990) Use of T7 RNA polymerase to direct expression of cloned genes. *Meth Enzymol* **185**, 60–89.
129. Studier FW, Daegelen P, Lenski RE, Maslov S and Kim JF. (2009) Understanding the differences between genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3) and comparison of the *E. coli* B and K-12 genomes. *J Mol Biol* **394**, 653–680.
130. Jeong H *et al.* (2009) Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J Mol Biol* **394**, 644–652 (2009).
131. van de Walle M and Shiloach J. (1998) Proposed mechanism of acetate accumulation in two recombinant *Escherichia coli* strains during high density fermentation. *Biotechnol Bioeng* **57**, 71–78.

132. Birney M, Um H and Klein C. (1997) Multiple levels of regulation of *Escherichia coli* succinyl-CoA synthetase. *Arch Biochem Biophys* **347**, 103–112.
133. Mann CJ, Mitchell T, *et al.* (1991) Phosphorylation and formation of hybrid enzyme species test the ‘half of sites’ reactivity of *Escherichia coli* succinyl-CoA synthetase. *Biochemistry* **30**, 1497–1503.
134. Joyce MA, Fraser, ME James M, *et al.* (2000) ADP-binding site of *Escherichia coli* succinyl-CoA synthetase revealed by X-ray crystallography. *Biochemistry* **39**, 17–25.
135. Joyce MA *et al.* (1999) Probing the nucleotide-binding site of *Escherichia coli* succinyl-CoA synthetase. *Biochemistry* **38**, 7273–7283.
136. Tretter L and Adam-Vizi V. (2000) Inhibition of Krebs cycle enzymes by hydrogen peroxide: A key role of [alpha]-ketoglutarate dehydrogenase in limiting NADH production under oxidative stress. *J Neurosci* **20**, 8972–8979.
137. Mailloux, RJ *et al.* (2007) The tricarboxylic acid cycle, an ancient metabolic network with a novel twist. *PLoS ONE* **2**, e690.
138. Ahn S, Jung J, Jang IA, Madsen EL and Park W. (2016) Role of glyoxylate shunt in oxidative stress response. *J Biol Chem* **291**, 11928–11938.
139. Rui B *et al.* (2010) A systematic investigation of *Escherichia coli* central carbon metabolism in response to superoxide stress. *BMC Syst Biol* **4**, 122.
140. Alhasawi A, Castonguay Z, Appanna ND, Auger C and Appanna VD. (2015) Glycine metabolism and anti-oxidative defence mechanisms in *Pseudomonas fluorescens*. *Microbiol Res* **171**, 26–31.
141. Messner KR and Imlay JA. (2002) Mechanism of superoxide and hydrogen peroxide formation by fumarate reductase, succinate dehydrogenase, and aspartate oxidase. *J Biol Chem* **277**, 42563–42571.
142. Gonidakis S, Finkel SE and Longo VD. (2010) Genome-wide screen identifies *Escherichia coli* TCA-cycle-related mutants with extended chronological lifespan dependent on acetate metabolism and the hypoxia-inducible transcription factor ArcA. *Aging Cell* **9**, 868–881.
143. Christodoulou D *et al.* (2018) Reserve flux capacity in the pentose phosphate pathway enables *Escherichia coli*'s rapid response to oxidative stress. *Cell Syst* **6**, 569–578.e7.
144. Brumaghim JL, Li Y, Henle E and Linn S. (2003) Effects of hydrogen peroxide upon nicotinamide nucleotide metabolism in *Escherichia coli*: changes in enzyme levels and nicotinamide nucleotide pools and studies of the oxidation of NAD(P)H by Fe(III). *J Biol Chem* **278**, 42495–42504.
145. Mamun AI *et al.* (2012) Identity and function of a large gene network underlying mutagenic repair of DNA breaks. *Science* **338**, 1344–1348.
146. Guelfo JR, Rodríguez-Rojas A, Matic I and Blázquez J. (2010) A MATE-family efflux pump rescues the *Escherichia coli* 8-oxoguanine-repair-



- deficient mutator phenotype and protects against H<sub>2</sub>O<sub>2</sub> killing. *PLoS Genet* **6**, e1000931.
147. Imlay, JA. (1995) A metabolic enzyme that rapidly produces superoxide, fumarate reductase of *Escherichia coli*. *J Biol Chem* **270**, 19767–19777.
  148. Seaver LC and Imlay JA. (2004) Are respiratory enzymes the primary sources of intracellular hydrogen peroxide? *J Biol Chem* **279**, 48742–48750.
  149. Fridovich, I. (1970) Quantitative aspects of the production of superoxide anion radical by milk xanthine oxidase. *J Biol Chem* **245**, 4053–4057.
  150. Altuvia S, Weinstein-Fischer D, Zhang A, et al. (1997) A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. *Cell* **90**, 43–53.
  151. Liochev SI and Fridovich I. (1992) Effects of overproduction of superoxide dismutases in *Escherichia coli* on inhibition of growth and on induction of glucose-6-phosphate dehydrogenase by paraquat. *Arch Biochem Biophys* **294**, 138–143.
  152. Luria SE and Delbruck M. (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511.
  153. Zheng, Q. (2002) Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation. *Math Biosci* **176**, 237–252.
  154. Zheng, Q. (2005) New algorithms for Luria-Delbrück fluctuation analysis. *Math Biosci* **196**, 198–214.
  155. Tajiri T, Maki H and Sekiguchi M. (1995) Functional cooperation of MutT, MutM and MutY proteins in preventing mutations caused by spontaneous oxidation of guanine nucleotide in *Escherichia coli*. *Mutat Res* **336**, 257–267.
  156. Bhatnagar SK and Bessman MJ. (1988) Studies on the mutator gene, *mutT* of *Escherichia coli*. Molecular cloning of the gene, purification of the gene product, and identification of a novel nucleoside triphosphatase. *J Biol Chem* **263**, 8953–8957.
  157. Sutherland P and McAlister-Henn L. (1985) Isolation and expression of the *Escherichia coli* gene encoding malate dehydrogenase. *J Bacteriol* **163**, 1074–1079.
  158. Lambert JM and Traut RR. (1981) The subunit interface of the *Escherichia coli* ribosome: Identification of proteins at the interface between the 30 S and 50 S subunits by crosslinking with 2-iminithiolane. *J Mol Biol* **149**, 451–476.
  159. Entine J and XiaoZhi L. (2015) Cheese: The GMO food die-hard GMO opponents love (and oppose a label for) Available from: <https://geneticliteracyproject.org/2015/05/15/cheese-gmo-food-die-hard-gmo-opponents-love-and-oppose-a-label-for/> Retrieved August 20, 2018.
  160. Jee, J. et al. (2016) Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* **534**, 693–696.
  161. Davis BD, Mingioli ES. (1950) Mutants of *Escherichia coli* requiring methionine or vitamin B12. *J Bacteriol* **60**, 17–28.

162. Blount ZD. (2016) A case study in evolutionary contingency. *Stud Hist Philos Biol Biomed Sci* **58**, 82–92.
163. Lenski RE, Travisano M. (1994) Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc Natl Acad Sci U S A* **91**, 6806–6814.
164. Lenski, R. E., Rose, M. R. & Simpson, S. C. (1991) Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *The American Naturalist*.
165. Rozen DE, Lenski RE. Long-term experimental evolution in *Escherichia coli*. VIII. (2000) Dynamics of a balanced polymorphism. *The American Naturalist* **155**, 24–35.
166. Le Gac M, Plucain J, Hindre T, Lenski RE, Schneider D. (2012) Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci U S A* **109**, 9487–9492.
167. Blount ZD, Borland CZ, Lenski RE. (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad Sci U S A* **105**, 7899–7906.
168. Quandt EM, Gollihar J, Blount ZD, Ellington AD, Georgiou G, Barrick JE. (2015) Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment. *eLife Sciences* **4**, e09696.
169. Quandt EM, Deatherage DE, Ellington AD, Georgiou G, Barrick JE. (2014) Recursive genomewide recombination and sequencing reveals a key refinement step in the evolution of a metabolic innovation in *Escherichia coli*. *Proc Natl Acad Sci U S A* **111**, 2217–2222.
170. Quandt, E. M. *et al.* Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment. *eLife Sciences* **4**, e09696 (2015).
171. Turner CB, Blount ZD, Mitchell DH, Lenski RE. (2015) Evolution and coexistence in response to a key innovation in a long-term evolution experiment with *Escherichia coli*; Preprint. Available from *bioRxiv*: 020958. Cited 25 January 2018.
172. Wiser MJ, Ribeck N, Lenski RE. (2013) Long-term dynamics of adaptation in asexual populations. *Science* **342**, 1364–1367.
173. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, *et al.* (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247.
174. Hall, BG. (1982) Chromosomal mutation for citrate utilization by *Escherichia coli* K-12. *J Bacteriol* **151**, 269–273.
175. van Hofwegen DJ, Hovde CJ, Minnich SA. (2016) Rapid evolution of citrate utilization by *Escherichia coli* by direct selection requires *citT* and *dctA*. *J Bacteriol* **198**, 1022–1034.

176. Roth JR, Maisnier-Patin S. (2016) Reinterpreting long-term evolution experiments: Is delayed adaptation an example of historical contingency or a consequence of intermittent selection? *J Bacteriol* **198**, 1009-1012.
177. Pos KM, Dimroth P, Bott, M. (1998) The *Escherichia coli* citrate carrier CitT: a member of a novel eubacterial transporter family related to the 2-oxoglutarate/malate translocator from spinach chloroplasts. *J Bacteriol* **180**, 4160–4165.
178. Lütgens M, Gottschalk G. (1980) Why a co-substrate is required for anaerobic growth of *Escherichia coli* on citrate. *J Gen Microbiol* **119**, 63–70.
179. Barrick JE, Lenski RE. (2009) Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harb Symp Quant Biol* **74**, 119–129.
180. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. (2017) The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50.
181. Raffaelli N, Lorenzi T, Mariani PL, Emanuelli M, Amic A, Ruggieri S *et al.* (1991) The *Escherichia coli* NadR regulator is endowed with nicotinamide mononucleotide adenylyltransferase activity. *J Bacteriol* **181**, 5509–5511.
182. Cortay JC, Negre D, Galinier A, Duclos B, Perriere G, Cozzone AJ. (1991) Regulation of the acetate operon in *Escherichia coli*: purification and functional characterization of the IclR repressor. *EMBO J* **10**, 675–679.
183. Perrenoud A, Sauer U. (2005) Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. *J Bacteriol* **187**, 3171–3179.
184. Salmon KA, Yang CR, Hatfield GW. (2013) Biosynthesis and regulation of the branched-chain amino acids. *EcoSal Plus* doi:10.1128/ecosalplus.3.6.1.5.
185. Lenski RE, Wiser MJ, Ribick N, Blount ZD, Nahum JR, Morris JJ, *et al.* (2015) Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *Proc Biol Sci* **282**, 20152292.
186. Fox JW, Lenski RE. (2015) From here to eternity—the theory and practice of a really long experiment. *PLoS Biol* **13**, e1002185.
187. Heard SB, Hauser DL. (1995) Key evolutionary innovations and their ecological mechanisms. *Taylor & Francis* **10**, 151–173.
188. Rohmer L, Hocquet D, Miller SI. (2011) Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol* **19**, 341–348.
189. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. (2016) Within-host evolution of bacterial pathogens. *Nat Rev Microbiol* **14**, 150–162.
190. Winter SE, Thiennimitr P, Winter MG, Butler BP, Huseby DL, Crawford RW, *et al.* (2010) Gut inflammation provides a respiratory electron acceptor for *Salmonella*. *Nature* **467**, 426–429.

191. Marvig RL, Damkiær S, Hossein Khademi SM, Markussen TM, Molin S, Jelsbak L. (2014) Within-host evolution of *Pseudomonas aeruginosa* reveals adaptation toward iron acquisition from hemoglobin. *MBio* **5** e00966–14.
192. Wiser MJ, Lenski RE. (2015) A comparison of methods to measure fitness in *Escherichia coli*. *PLoS ONE* **10** e0126210–11.
193. Lenski RE. (2003) Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*, in Plant Breeding Reviews: Long-term Selection: Crops, Animals, and Bacteria, Volume 24, Part 2 (ed J. Janick), John Wiley & Sons, Inc., Oxford, UK
194. Stead MB, Agrawal A, Bowden KE, Nasir R, Mohanty BK, Meagher RB *et al.* (2012) RNA snap <sup>TM</sup> : a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. *Nucleic Acids Res* **40**, e156.
195. Barrick JE, Colburn G, Deatherage DE, Traverse CC, Strand MD, Borges JJ, *et al.* (2014) Identifying structural variation in haploid microbial genomes from short-read resequencing data using *breseq*. *BMC Genomics* **15**, 1039.
196. Feisenstein, J. (1989) PHYLIP- Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166.
197. Keulartz J, van den Belt H. (2016) DIY-Bio – economic, epistemological and ethical implications and ambivalences. *Life Sci Soc Pol* **12**, 7.
198. Callaway E. (2015) Lab staple agar runs low. *Nature* **528**, 171–172.
199. Giavasis I, Harvey LM, McNeil B. (2000) Gellan gum. *Crit Rev Biotechnol* **20**, 177–211.
200. Pollock TJ. (1993) Gellan-related polysaccharides and the genus *Sphingomonas*. *J Gen Microbiol* **139**, 1939–1945.
201. Fialho AM, Moreira LM, Granja AT, Popescu AO, Hoffmann K, Sá-Correia I. (2008) Occurrence, production, and applications of gellan: current state and perspectives. *Appl Microbiol Biotechnol* **79**, 889–900.
202. Nampoothiri KM, Singhanian RR, Sabarinath C, Pandey A. (2003) Fermentative production of gellan using *Sphingomonas paucimobilis*. *Process Biochem* **38**, 1513–1519.
203. Park SF. (2012) Microbiology at home: a short non-laboratory manual for enthusiasts and bioartists. Available from: <https://exploringtheinvisible.files.wordpress.com/2013/11/manual2013.pdf>. Retrieved April 3, 2018.
204. Emmert EAB and ASM Task Committee on Laboratory Biosafety. (2012) ASM guidelines for biosafety in teaching laboratories. Available from: [www.asm.org/images/asm\\_biosafety\\_guidelines-FINAL.pdf](http://www.asm.org/images/asm_biosafety_guidelines-FINAL.pdf). Retrieved April 3, 2018.
205. Wu X, Wu R, Li O, Zhu L, Chen Y, Qian C, Chen M. (2014) Yellow pigments generation deficient *Sphingomonas* strain and application thereof in gellan gum production. *US patent* 8,685,698 B2.

206. Kang KS, Veeder GT, Colegrove GT. (1983) Deacetylated polysaccharide S-60. *US patent* 4,385,123.